

Wirth, Joachim; Lebens, Morena Individualdiagnostik

2011, 36 S.



Quellenangabe/ Reference:

Wirth, Joachim; Lebens, Morena: Individualdiagnostik. 2011, 36 S. - URN:
urn:nbn:de:0111-pedocs-107485 - DOI: 10.25656/01:10748

<https://nbn-resolving.org/urn:nbn:de:0111-pedocs-107485>

<https://doi.org/10.25656/01:10748>

Nutzungsbedingungen

Gewährt wird ein nicht exklusives, nicht übertragbares, persönliches und beschränktes Recht auf Nutzung dieses Dokuments. Dieses Dokument ist ausschließlich für den persönlichen, nicht-kommerziellen Gebrauch bestimmt. Die Nutzung stellt keine Übertragung des Eigentumsrechts an diesem Dokument dar und gilt vorbehaltlich der folgenden Einschränkungen: Auf sämtlichen Kopien dieses Dokuments müssen alle Urheberrechtshinweise und sonstigen Hinweise auf gesetzlichen Schutz beibehalten werden. Sie dürfen dieses Dokument nicht in irgendeiner Weise abändern, noch dürfen Sie dieses Dokument für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, aufführen, vertreiben oder anderweitig nutzen.

Mit der Verwendung dieses Dokuments erkennen Sie die Nutzungsbedingungen an.

Terms of use

We grant a non-exclusive, non-transferable, individual and limited right to using this document.

This document is solely intended for your personal, non-commercial use. Use of this document does not include any transfer of property rights and it is conditional to the following limitations: All of the copies of this documents must retain all copyright information and other information regarding legal protection. You are not allowed to alter this document in any way, to copy it for public or commercial purposes, to exhibit the document in public, to perform, distribute or otherwise use the document in public.

By using this particular document, you accept the above-stated conditions of use.

Kontakt / Contact:

peDOCS
DIPF | Leibniz-Institut für Bildungsforschung und Bildungsinformation
Informationszentrum (IZ) Bildung
E-Mail: pedocs@dipf.de
Internet: www.pedocs.de

Mitglied der


Leibniz-Gemeinschaft

Individualdiagnostik

Joachim Wirth
Morena Lebens



UDiKom

**Aus- und Fortbildung der Lehrkräfte
in Hinblick auf Verbesserung der
Diagnosefähigkeit, Umgang mit
Heterogenität, individuelle Förderung**

Die Produktion dieses Materials
zum Einsatz in die Lehrerbildung
wurde ermöglicht durch

Deutsche Telekom Stiftung



Individualdiagnostik

Alle im Projekt erstellten Materialien
finden Sie unter
www.udikom.de



Inhaltsverzeichnis – Teil 1 – Individualdiagnostik

1.1	Gegenstand und Zielsetzung	3
1.1.1	Weiterführende Literatur	4
1.2	Bezugsnormen	5
1.2.1	Kriteriale Bezugsnorm	5
1.2.2	Soziale Bezugsnorm	5
1.2.3	Individuelle Bezugsnorm	6
1.2.4	Bezugsnorm im Vergleich	7
1.2.5	Weiterführende Literatur	7
1.3	Testkonstruktion	8
1.3.1	Individualdiagnostik mit Hilfe von Tests	8
1.3.2	Das Testergebnis	9
1.3.3	Indikatorbildung	10
1.3.4	Kriterien der Qualität von Tests	11
1.3.4.1	Reliabilität	11
1.3.4.2	Validität	14
1.3.4.3	Objektivität	17
1.3.5	Nebengütekriterien	18
1.3.5.1	Normierung	19
1.3.6	Weiterführende Literatur	19
1.3.7	Verständnis- und Diskussionspunkte	19
1.4	Inhaltlicher Anwendungsbereich/Phänomenbereich	20
1.4.1	Schulleistungsmerkmale	21
1.4.1.1	Schulleistungstests	21
1.4.1.2	Hamburger Schulleistungstest für vierte und fünfte Klassen	21
1.4.1.3	Was ist ein „guter“ Schulleistungstest?	22
1.4.2	Schulleistungsrelevante Merkmale: Intelligenz	22
1.4.2.1	Eindimensionale Intelligenztests: Der Hamburg-Wechsler-Intelligenz-Test	23
1.4.2.2	Mehrdimensionale Intelligenztests: Das Leistungsprüfsystem	24
1.4.2.3	Sprachfreie Tests: Der Culture Fair Test	25
1.4.3	Schulleistungsrelevante Merkmale: Motivation	25
1.4.3.1	Erfassung der Lern- und Leistungsmotivation: Der SELLMO	25
1.4.4	Schulleistungsrelevante Merkmale: Fähigkeitsselbstkonzept	26
1.4.4.1	Erfassung des schulischen Selbstkonzepts: Der SESSKO	27
1.4.5	Weiterführende Literatur	27
1.4.6	Verständnis- und Diskussionspunkte	27
1.5	Praktische Implikationen	28
1.5.1	Validität	28
1.5.2	Objektivität	29
1.5.2.1	Schriftliche Testaufgaben mit geschlossenem Antwortformat	29
1.5.2.2	Schriftliche Testaufgaben mit offenem Antwortformat	32
1.5.3	Reliabilität	33
1.5.4	Weiterführende Literatur	33
1.6	Literatur	34

1.1 Gegenstand und Zielsetzung

In diesem Kapitel behandelte Fragen:

- *Welchen pädagogischen Ertrag bieten individualdiagnostische Verfahren?*
- *Wodurch zeichnet sich die pädagogische Individualdiagnostik aus?*

Eine der wichtigsten Aufgaben von Lehrern ist sicherlich die genaue Einschätzung ihrer Schüler, sei es im Bezug auf ihre Fachleistungen, ihre Lernmotivation, ihre sozialen Fähigkeiten oder andere Merkmale, die Lernen beeinflussen können. Diagnostik ist daher ein bedeutsamer Teil der alltäglichen Arbeit im Klassenzimmer. Durch die richtige Einschätzung der Leistungen und Fähigkeiten der Schüler können diese optimal gefördert werden, was wiederum einen enormen Einfluss auf die Motivation der Schüler hat. Falsche Entscheidungen können dagegen die Bildungskarriere eines Schülers sehr erschweren. Darum ist es sehr wichtig, dass pädagogische Diagnostik niemals aus dem Bauch heraus betrieben wird, sondern immer systematisch, geplant und objektiv ist.

Hintergrund

Gute pädagogische Diagnostik ist jedoch nicht nur entscheidend für den Schüler. Auch Lehrer profitieren, wenn ihre pädagogischen Entscheidungen auf solider Diagnostik basieren: Jeder im aktiven Schuldienst kennt bspw. die Situation, wenn ein Schüler auf Grund schlechter Zeugnisnoten nicht versetzt wird. Beim Gespräch mit den Eltern sind sich diese dann sicher: Die 5 in Deutsch ist völlig unberechtigt, nicht der Schüler ist zu schlecht, sondern die Klassenarbeiten waren viel zu schwer und überhaupt; im Unterricht wurden doch ganz andere Dinge behandelt als die, die hinterher abgefragt wurden; Schuld ist der Lehrer, der den Schüler doch schon seit dem letzten Schuljahr immer strenger bewertet als den Rest der Klasse. Eine schwierige Situation für jeden Lehrer. Umso wichtiger sind hier gute Argumente dafür, dass der Lehrer den Schüler richtig bewertet hat. Genau diese Argumente liefert eine gute pädagogische Individualdiagnostik.

Ziel der Individualdiagnostik ist es, die Ausprägung verschiedener psychologischer Merkmale genau zu erfassen. Für Lehrer bedeutet dies: durch die Individualdiagnostik lässt sich bspw. einschätzen, wie gut oder schlecht die Leistungen und Fähigkeiten einzelner Schüler sind. Aber es geht nicht nur um Leistung. Gute Individualdiagnostik kann bspw. auch Auskunft darüber geben, wie Schüler selbst ihre Fähigkeiten einschätzen, wie motiviert sie sind, ob sie ängstlich sind oder vieles mehr.

Ziel der
Individual-
diagnostik

Die Individualdiagnostik kann somit Informationen liefern, die Lehrkräfte benötigen, um bildungsbezogene Entscheidungen für die jeweilige Schülerin/den jeweiligen Schüler begründet treffen zu können. Soll der Schüler versetzt werden? Welche Form der Unterstützung braucht der Jugendliche? Bleibt der Schüler aufgrund falscher Selbsteinschätzung hinter seinen Möglichkeiten zurück? Solche Fragen lassen sich durch eine gute Individualdiagnostik objektiv beantworten, und die daraus gezogenen Konsequenzen lassen sich nachvollziehbar begründen.

In Unterrichtssituationen treffen Lehrkräfte pädagogische Entscheidungen häufig auf Grundlage zufälliger sowie auch geplanter Beobachtungen der Schülerinnen und Schüler (Kliemann, 2008). Für eine Vielzahl pädagogischer Entscheidungen sind solche Beobachtungen auch ausreichend, zumal sie sehr ökonomisch, bspw. während des Unterrichts ohne nennenswerten zusätzlichen Aufwand, durchgeführt werden können. Allerdings müssen sich Lehrkräfte auch darüber bewusst sein, dass derartige Beobachtungen die tatsächlichen Fähigkeiten von Schülerinnen und Schülern häufig nur ungenau abbilden. Daher geben Schrader und Helmke (2001) zu bedenken: „Eine zutreffende Einschätzung des Leistungsstandes ist allerdings eine außerordentlich schwierige Aufgabe, die ohne den Einsatz von professionell entwickelten, am Lehrplan orientierten diagnostischen Instrumenten kaum möglich ist“ (S. 50).

Individualdiagnostische Verfahren sind daher eine notwendige Ergänzung zu Lehrerbeobachtungen. Dabei kann die Individualdiagnostik verschiedene Funktionen erfüllen. Bspw. können individualdiagnostische Verfahren als Lernausgangsdia gnose Auskunft über den Vorwissensstand oder die Motivation der Schülerinnen und Schüler bieten. Prozessbegleitend kann die Individualdiagnostik innerhalb einer Unterrichtssequenz zur Nachvollziehbarkeit individueller Lernprozesse beitragen. Am Ende der Unterrichtssequenz dient die Individualdiagnostik der Erfassung und Analyse des Lernergebnisses.

Funktionen
der
Individual-
diagnostik

Im vorliegenden Studienbriefteil 1 liegt das Hauptaugenmerk auf der Diagnose von Schulleistung und den Merkmalen, die Schulleistung beeinflussen können (schulleistungsrelevante Merkmale). Dabei richtet sich die Diagnose auf Merkmale einzelner Personen. Für die Diagnose von Schulleistung und schulleistungsrelevanten Merkmalen stehen verschiedene individualdiagnostische Verfahren (Testinstrumente) zur Verfügung, die in diesem Studienbrief exemplarisch vorgestellt werden.

Alle individualdiagnostischen Verfahren haben gemein, dass sie am Ende ein Ergebnis in Form einer Zahl liefern. Bspw. könnte das Ergebnis in einem standardisierten Lesegeschwindigkeitstest 837 lauten. Diese Zahl ist zunächst einmal nicht aussagekräftig und zwar aus zwei Gründen: Zum einen ist sie nur durch einen Vergleich mit einer sogenannten „Bezugsnorm“ sinnvoll interpretierbar. Zum anderen macht es nur Sinn, diese Zahl überhaupt zu interpretieren, wenn der eingesetzte Lesegeschwindigkeitstest bestimmten Qualitätskriterien genügt. Es ist daher unerlässlich, dass Lehrkräfte sowohl die verschiedenen Bezugsnormen und ihre Bedeutung kennen als auch dazu in der Lage sind, diagnostische Verfahren in Bezug auf verschiedene Qualitätskriterien zu überprüfen. In Kapitel 1.2 dieses Stu-

Testtheorie –
Wozu?

dienbriefs können Sie sich über die verschiedenen Bezugsnormen und ihre Bedeutung informieren. Kapitel 1.3 stellt Ihnen dann die verschiedenen Qualitätskriterien für Testverfahren (die sog. „Testgütekriterien“) vor.

Die in diesen beiden Grundlagenkapiteln zu erwerbenden testtheoretischen Kenntnisse über Bezugsnormen und Qualitätskriterien sind zum einen die Voraussetzung dafür, dass Lehrkräfte dazu in der Lage sind, aus dem Angebot existierender Testverfahren das für ihre Zwecke am besten geeignete herauszusuchen (Eine Datenbank, in der für Lehrkräfte interessante Testverfahren gelistet sind, finden Sie unter <http://tests.udikom.de/>). Zum anderen sind diese Kenntnisse notwendig, um die Aussagekraft individualdiagnostischer Ergebnisse, die bspw. durch den schulpyschologischen Dienst erfasst wurden, nachzuvollziehen und entsprechende pädagogische Entscheidungen ableiten zu können. Doch nicht nur die Auswahl und Interpretation bestehender individualdiagnostischer Testverfahren erfordern diese Grundlagenkenntnisse. Auch die selbstständige Entwicklung von Testverfahren – ein alltägliches Geschäft von Lehrkräften, wenn sie bspw. Klassenarbeiten konzipieren – sollte von diesen Grundkenntnissen geleitet sein. Das bedeutet nicht, dass Lehrkräfte jede ihrer Klassenarbeiten nach dem Vorbild etablierter und von wissenschaftlichen Verlagen publizierter individualdiagnostischer Testverfahren entwickeln und prüfen müssen. Dass dieser Aufwand nicht bei jeder Messung im schulischen Alltag von Lehrerinnen und Lehrern, wie z.B. bei Klassenarbeiten geleistet werden kann, ist selbstverständlich. Trotzdem sind diese testtheoretischen Kenntnisse hilfreich, um die Qualität auch von Diagnoseinstrumenten wie Klassenarbeiten zu verbessern. Wie es Lehrkräften möglich ist, das diagnostische Potenzial von Klassenarbeiten zu erhöhen, ist Gegenstand des letzten Kapitels dieses Studienbriefs.

Zusammen-
fassung

Individualdiagnostische Verfahren können in der pädagogischen Praxis zur Diagnose von Schulleistung und schulleistungsrelevanten Merkmalen zu verschiedenen Zeiten eingesetzt werden. Die Ergebnisse dienen als Entscheidungsgrundlage für bildungsbezogene Entscheidungen im Einzelfall. Zur Auswahl und Bewertung verschiedener individualpsychologischer Verfahren, zur Interpretation der Ergebnisse sowie für die selbstständige Entwicklung von Tests wie z.B. Klassenarbeiten sind Kenntnisse über verschiedene Bezugsnormen sowie über verschiedene Qualitätskriterien für Testverfahren notwendig. Der vorliegende Studienbrief möchte genau diese Kenntnisse und Fähigkeiten vermitteln.

1.1.1 Weiterführende Literatur

- Paradies, L., Linser, H.J., & Greving, J. (2007). *Diagnostizieren, Fordern und Fördern*. Berlin: Cornelsen Verlag Scriptor.
- Fisseni, H.-J. (2004). *Lehrbuch der Psychologischen Diagnostik* (3. Aufl.). Göttingen: Hogrefe.
- Ingenkamp, K. (1997). *Lehrbuch der pädagogischen Diagnostik*. Weinheim: Beltz.
- Schweizer, K. (2006). *Leistung und Leistungsdiagnostik*. Heidelberg: Springer.

1.2 Bezugsnormen

Fragen, die in diesem Kapitel beantwortet werden:

- Was sind Bezugsnormen?
- Welchen Beitrag leisten Bezugsnormen für die Interpretation und Bewertung von Testergebnissen?
- Was ist bei der Nutzung der jeweiligen Bezugsnorm zu berücksichtigen?

Wie in Kapitel 1.1 bereits erwähnt, haben alle Testverfahren gemeinsam, dass sie am Ende ein Ergebnis in Form einer Zahl liefern. Die interessierende Ausprägung einer Personeneigenschaft wird durch ein numerisches Testergebnis ausgedrückt. Daraus ergeben sich Aussagen wie z.B. „Die Lesegeschwindigkeit von Florian ist 837“ oder „Annas Testängstlichkeit liegt bei 3,6“. Diese Aussagen sind zunächst einmal nichtssagend. Sie enthalten zwar das numerische Testergebnis, den sogenannten „Rohwert“ (im Falle von Florian 837, bei Anna 3,6). Dieser Rohwert ist jedoch ohne weitere Informationen inhaltlich nicht interpretierbar oder bewertbar. Ist eine Lesegeschwindigkeit von 837 als gut oder schlecht zu bewerten? Schneiden Schülerinnen und Schüler, die älter als Florian sind, im Lesegeschwindigkeitstest besser als Florian ab? Ist eine Testängstlichkeit von 3,6 normal? Kann Anna mit Testsituationen jetzt besser umgehen als noch vor einem Jahr? Fragen, die für eine Interpretation und Bewertung des Testergebnisses bedeutsam sind, die aber aufgrund des Rohwerts allein nicht beantwortet werden können.

Rohwert

Die Interpretation und Bewertung eines Rohwerts erfolgt über einen oder mehrere Vergleiche. Dabei wird der Rohwert in Bezug gesetzt zu einer weiteren Zahl. Diese weitere Zahl dient als Norm, durch den Vergleich mit ihr wird der Rohwert normiert. Diese Norm, die in Form einer konkreten Zahl oder auch als Verteilung von Zahlen vorliegen kann, wird Bezugsnorm genannt. Sie kann als Standard oder Maßstab angesehen werden, anhand dessen das Testergebnis (und damit die Ausprägung der interessierenden Personeneigenschaft) beurteilt wird.

Normierung
des Rohwerts

Üblicherweise unterscheidet man drei verschiedene Arten von Bezugsnormen, die in Bezug auf unterschiedliche Fragestellungen der Individualdiagnostik Anwendung finden. Die *kriteriale Bezugsnorm* gibt an, wie hoch das Testergebnis einer Person, der Rohwert, mindestens sein muss, damit das Testergebnis positiv bewertet werden kann. Mit Hilfe einer *sozialen Bezugsnorm* wird das Testergebnis einer Person mit den entsprechenden Ergebnissen vergleichbarer Personen verglichen. Verwendet man eine *individuelle Bezugsnorm*, dann vergleicht man das Testergebnis einer Person mit Ergebnissen, die dieselbe Person zu früheren Zeitpunkten bereits in demselben Test erzielen konnte. Im Folgenden werden diese drei Bezugsnormen näher beleuchtet.

Bezugs-
normen

1.2.1 Kriteriale Bezugsnorm

Die kriteriale Bezugsnorm bezieht sich immer auf ein festes Kriterium, im Unterricht ist dies das Lehrziel, das sich ein Lehrer gesetzt hat. Unter Verwendung einer kriterialen Bezugsnorm wird bspw. überprüft, ob die Leistung eines Schülers einen bestimmten Standard erreicht oder nicht. Die kriteriale Bezugsnorm kommt auch insbesondere dann zum Einsatz, wenn der Leistungsbeurteilung eine qualifizierende oder berechtigende Funktion zukommt und somit bestimmte Standards erreicht werden müssen (Rheinberg, 2001).

Beispiel: Kriteriale Bezugsnorm

Denken wir uns einen Lehrer, der mit seiner Klasse ein Lesetraining durchgeführt hat mit dem Ziel, dass alle Schüler der Klasse einen Text flüssig und mit einer bestimmten Geschwindigkeit lesen können. Um zu überprüfen, ob er dieses Ziel erreicht hat, führt der Lehrer am Ende des Trainings einen standardisierten Lesegeschwindigkeitstest durch. Bei diesem Test wird ein Lesetext vorgegeben, den Schüler innerhalb von vier Minuten lesen sollen. Vorab definiert der Lehrer als Kriterium, dass die Schüler in der vorgegebenen Zeit mindestens 1000 Wörter verstehend lesen können. Damit dient die Zahl 1000 als kriteriale Bezugsnorm. Der Lehrer wird Schülern, die 1000 oder mehr Wörter in der gegebenen Zeit gelesen haben, eine ausreichende Lesegeschwindigkeit bescheinigen und den anderen Schülern entsprechend nicht.

1.2.2 Soziale Bezugsnorm

Bei der sozialen Bezugsnorm dienen als Vergleichswerte die Testergebnisse anderer vergleichbarer Personen. Die Beurteilung eines einzelnen Testergebnisses ist somit davon abhängig, wie hoch oder niedrig die Testergebnisse dieser vergleichbaren Personen ausfallen. Die Nutzung einer sozialen Bezugsnorm ist dann sinnvoll, wenn es bspw. Ziel der Leistungsbeurteilung ist, den Besten oder die Beste aus einer Gruppe zu ermitteln, wie man es z.B. von sportlichen Wettkämpfen kennt. Wie Rheinberg (2001) unterstreicht, ist eine Beurteilung anhand der sozialen Bezugsnorm jedoch nicht notwendigerweise auf Selektion und Auslese der Besten beschränkt, sondern kann auch eine Grundlage für gezielte Fördermaßnahmen bieten, bspw. wenn für bestimmte Förderangebote eine nur begrenzte Anzahl von Plätzen verfügbar ist und diese Plätze denjenigen gegeben werden sollen, die sie am meisten benötigen.

Beispiel: Soziale Bezugsnorm

Nehmen wir an, ein Schüler hätte im Lesegeschwindigkeitstest innerhalb von vier Minuten 922 Wörter gelesen. Er hätte damit das Kriterium des Lehrers von 1000 Wörtern (s. Beispiel in Kap. 1.2.1) nicht erreicht und bekäme beim Anlegen einer kriterialen Bezugsnorm eine entsprechend schlechte Bewertung. Was aber, wenn eine soziale Bezugsnorm angelegt wird? In dem Fall könnte er durchaus eine sehr gute Bewertung erhalten, nämlich genau dann, wenn alle oder zumindest viele seiner Mitschüler weniger als 922 Wörter gelesen haben.

Normierung des Testverfahrens Viele individualdiagnostische Testverfahren, die meist über Testverlage vertrieben werden, haben die Eigenschaft, dass sie „normiert“ sind (s.a. Kap. 1.3). Das bedeutet, dass im Begleitheft zu den Tests sog. Normtabellen enthalten sind, die Vergleichswerte umfangreicher Stichproben präsentieren. Diese Stichproben, die häufig mehrere tausend Personen umfassen, repräsentieren bestimmte Populationen. So findet man häufig separate Normtabellen für Männer und Frauen, Normtabellen für unterschiedliche Altersklassen oder Klassenstufen, etc. Anhand dieser Normtabellen kann jeder Testanwender ein individuelles Testergebnis in Bezug auf vergleichbare Populationen einordnen, ohne dass er selbst Vergleichswerte erheben muss.

Schwächen im schulischen Kontext Wird die soziale Bezugsnorm im schulischen Kontext angewendet, so sollten auch deren Schwachpunkte beachtet werden. Rheinberg (2001) weist auf zwei wesentliche blinde Flecken hin:

- Die soziale Bezugsnorm wird häufig auf ein klasseninternes Bezugssystem reduziert. Gilt es bspw. die Leistung eines Schülers zu bewerten, so erfolgt die Leistungsbeurteilung oftmals im Vergleich zu den entsprechenden Leistungen des jeweiligen Klassenverbands. Die Lehrkraft vergleicht innerhalb einer Klasse oder auch innerhalb einer Jahrgangsstufe. Eine einzelne Klasse oder Jahrgangsstufe ist jedoch nicht repräsentativ für eine ganze Population, und ihre durchschnittliche Leistung kann beträchtlich über oder auch unter der durchschnittlichen Leistung der Population liegen. Die Beurteilung einer einzelnen Schülerleistung hängt demnach in besonders starkem Maß davon ab, ob der Schüler in einer (verhältnismäßig) leistungsstarken oder leistungsschwachen Klasse ist. Dadurch entsteht so etwas wie der so genannte „*big fish – little pond*“-Effekt (Fischteicheffekt): Mittelstarke Schülerinnen und Schüler sind in leistungsschwachen Klassen die großen Fische im kleinen Teich, das bedeutet, dass sie – bei gleicher Leistung – besser bewertet werden als Schülerinnen und Schüler in leistungsstarken Klassen.
- Sowohl eine positive als auch eine negative Leistungsentwicklung des gesamten Klassenverbands wird bei Anwendung der sozialen Bezugsnorm ausgeblendet, da nur die Leistungsunterschiede zwischen den Schülerinnen und Schülern zählen. Das geht einher mit dem empirischen Befund, dass über 50 % der Schülerinnen und Schüler von Lehrkräften, deren Leistungsbeurteilung auf einer sozialen Bezugsnorm fußt, keine Leistungssteigerung über das Schuljahr hinweg erkennen konnten oder sogar davon ausgehen, am Ende des Schuljahres weniger zu können als am Anfang (Rheinberg, 1980).

1.2.3 Individuelle Bezugsnorm

Die individuelle Bezugsnorm ergibt sich aus den Testergebnissen einer Person, die diese zu früheren Zeitpunkten in demselben Test oder in parallelen Tests (vgl. Kap 1.3.3.1) erzielen konnte. Dadurch wird die Entwicklung eines Schülers abbildbar. Konnte sich ein Schüler in seiner Leistung steigern? Hat er etwas dazu gelernt? Hat eine Unterrichtseinheit das Interesse an dem behandelten Thema wecken oder erhöhen können? Sollen Fragen dieser Art beantwortet werden, so ist eine individuelle Bezugsnorm heranzuziehen.

Beispiel: Individuelle Bezugsnorm

Nehmen wir an, der Lehrer aus dem Beispiel in Kap. 1.2.1 hätte den Lesegeschwindigkeitstest nicht nur nach dem Lesetraining eingesetzt, sondern in einer Parallelversion auch davor. Sein Ziel sei es, die Lesegeschwindigkeit seiner Schüler zu erhöhen. Wenn seine Schüler vor dem Training im Durchschnitt 850 Wörter gelesen hätten, dann könnte der Lehrer zufrieden sein, wenn seine Schüler nach dem Training im Durchschnitt 910 Wörter lesen könnten, da sie nach dem Training im Durchschnitt 60 Wörter mehr innerhalb der vier Minuten lesen konnten als noch vor dem Training.

Vor- und Nachteile Das Anlegen einer individuellen Bezugsnorm kann gerade bei leistungsschwächeren Schülerinnen und Schülern enorme motivationale Vorteile bieten. Auch wenn sie in einem Test, gemäß einer kriterialen Bezugsnorm, eine eher geringe Leistung zeigen, so kann es doch sehr motivierend sein, wenn es Beachtung findet, falls diese (geringe) Leistung wenigstens besser ist als eine entsprechende Leistung zu einem früheren Zeitpunkt. Wenn man Schülerinnen und Schülern eine solche positive Entwicklung zurückmelden kann, kann dies einen enormen Effekt auf deren Motivation haben (Rheinberg, 1980).

Selbstredend hätte eine ausschließliche Beschränkung auf eine individuelle Bezugsnorm jedoch auch bizarre Folgen, bspw. wenn eine Leistungsbeurteilung eine berechtigende Funktion inne hat (z.B. Zeugnisnoten). Bei konsequenter

Anwendung der individuellen Bezugsnorm würde ein Schüler, der zu Beginn eines Schuljahres eine schlechte Leistung zeigte, sich aber zum Ende des Schuljahres auf eine wenigstens durchschnittliche Leistung steigern konnte, eine bessere Zeugnisnote erhalten als ein Schüler, der von Beginn an eine durchschnittliche Leistung zeigte. Die individuelle Bezugsnorm ist also hilfreich, wenn es darum geht, die Entwicklung von Schülerinnen und Schülern abzubilden, und diese Fortschritte den Schülerinnen und Schülern auch zurückmelden zu können. Wenn der Leistungsbeurteilung jedoch eine berechtigende Funktion zukommt, dann sollte für eine solche pädagogische Entscheidung eher eine kriteriale oder eine soziale Bezugsnorm genutzt werden.

1.2.4 Bezugsnorm im Vergleich

	Kriteriale Bezugsnorm	Soziale Bezugsnorm	Individuelle Bezugsnorm
Vorteile	Bewertung unabhängig von (1) sozialen Vergleichen und von (2) der individuellen Leistungssteigerung	Ermöglicht soziale Vergleiche mit einer Bezugsgruppe	„Schwankungen im Lernverlauf werden unter individueller Bezugsnorm wie unter einem Vergrößerungsglas sichtbar gemacht“ (Rheinberg, 2001)
Nachteile	Nicht auf die Erfassung individueller Lernfortschritte ausgerichtet	Klasseninternes Bezugssystem: Big fish – little pond-Effekt Ausbblendung von Leistungsschwankungen im Klassenverband	Selbstbeurteilung mittels sozialer Vergleiche nicht möglich

Tabelle 1: Vor- und Nachteile der verschiedenen Bezugsnormen

Jede der Bezugsnormen hat ihre spezifischen Vor- und Nachteile (Tabelle 1). Welche Bezugsnorm für die Interpretation eines Testwertes die richtige ist, hängt von der Art der pädagogischen Entscheidung ab.

Die individuelle Bezugsnorm ist von besonderer Bedeutung, wenn man sich mit der Entwicklung von Lernenden über einen bestimmten Zeitraum hinweg beschäftigt. So können Lernfortschritte einzelner Schülerinnen und Schüler über ein Schuljahr hinweg untersucht oder auch die Effekte pädagogischer Interventionsmaßnahmen geprüft werden. Die kriteriale Bezugsnorm kommt zum Einsatz, wenn ermittelt werden soll, inwieweit einzelne Schülerinnen und Schüler bestimmte curriculare Standards erfüllen. Die soziale Bezugsnorm wird herangezogen, um zu bilanzieren, inwiefern Schülerinnen und Schüler hinsichtlich ihrer kognitiven, affektiv-motivationalen oder beider Merkmale im Vergleich zu einer entsprechenden Bezugsgruppe mit vergleichbaren Eingangsvoraussetzungen abweichen. Auf die soziale und kriteriale Bezugsnorm wird im Studienbrief „Vergleichsarbeiten“ verstärkt eingegangen.

Die Bezugsnormen sind keinesfalls auf den pädagogisch-diagnostischen Bereich beschränkt, sondern finden sich häufig auch im Alltag wieder (Rheinberg, 2001). Nehmen wir als Beispiel die derzeit populären Castingshows, in denen der oder die beste Sängerin gekürt werden soll. Seitens der Jury oder der Moderatoren werden oft Äußerungen vorgenommen wie: „Von allen Kandidaten hat sie eindeutig die beste Performance gezeigt“ (soziale Bezugsnorm), „Er ist heute über sich hinausgewachsen. Diese Steigerung hätte in den letzten Shows niemand für möglich gehalten - eine starke Leistung“ (individuelle Bezugsnorm) oder „Sie hat eine glockenklare Stimme und hat perfekt intoniert“ (kriteriale Bezugsnorm). Nur ein Beispiel, das zeigt, dass wir in unserem Alltag, bewusst oder unbewusst, ständig auf Bezugsnormen zurückgreifen, auch wenn diese oftmals nicht die psychometrische Qualität wie bspw. die Normtabelle in standardisierten Testverfahren haben. Doch unabhängig davon ist die Verwendung einer der drei vorgestellten Bezugsnormen ein alltägliches Geschäft.

1.2.5 Weiterführende Literatur

Sacher, W. (2009). *Leistungen entwickeln, überprüfen und beurteilen. Bewährte und neue Wege für die Primar- und Sekundarstufe*. Bad Heilbrunn: Klinkhardt.

Winter, F. (2008). *Leistungsbewertung*. Hohengehren: Schneider-Verlag.

1.3 Testkonstruktion

Gegenstand dieses Kapitels ist zum einen der prinzipielle Aufbau individualdiagnostischer Testverfahren. Zudem werden Kriterien besprochen, anhand derer die Qualität von Testverfahren beurteilt werden kann. Dafür werden testtheoretische Kenntnisse vermittelt, die für ein Verständnis des Aufbaus individualdiagnostischer Verfahren sowie für ihre Bewertung unverzichtbar sind. Zu diesem Zwecke werden in diesem Kapitel folgende Fragen behandelt:

- Was genau ist ein Test?
- Welche Eigenschaften hat ein Testergebnis?
- Was zeichnet ein qualitativ hochwertiges individualdiagnostisches Testverfahren aus?
- Worauf soll bei der Auswahl eines Testinstruments geachtet werden?

1.3.1 Individualdiagnostik mit Hilfe von Tests

Was ist ein Test?

Tests, die einem Auskunft über bestimmte Eigenschaften einer Person (meist von sich selbst) versprechen, finden sich zu Hauf im Internet oder in Zeitschriften. Zehn plausibel klingende Fragen beantworten, und schon weiß man, wo die eigenen Stärken und Schwächen liegen, ob man über Sozialkompetenz verfügt, ob man ehrgeizig, einfühlsam oder auch belastbar ist, ob man gut zuhören kann oder ähnliches. Der gesunde Menschenverstand sagt einem aber bereits, dass ein „Test“, der auf einer der hinteren Seiten einer TV-Zeitschrift abgedruckt ist, qualitativ wahrscheinlich nicht vergleichbar ist mit Testverfahren, die wissenschaftlich fundiert entwickelt und auf ihre Qualität hin überprüft wurden. Die Frage ist jedoch, worin genau der Unterschied in der Qualität liegt? Anhand welcher Kriterien lassen sich qualitativ hochwertige von qualitativ minderwertigen Testverfahren unterscheiden? Was sind die Voraussetzungen, die bei der Entwicklung und Überprüfung von individualdiagnostischen Tests gegeben sein müssen, damit ein qualitativ guter Test entstehen kann? Und nicht zuletzt: Was genau ist eigentlich ein Test?

Lienert (1961) betont vier Merkmale, die einen Test ausmachen, und die eine notwendige, wenn auch nicht hinreichende Voraussetzung für eine hohe Testqualität sind:

Definition: Test

„Nicht jede, zu diagnostischen Zwecken aufgestellte Untersuchung kann als Test gelten, sondern nur eine solche, die

- erstens wissenschaftlich begründet ist,
- zweitens routinemäßig – also unter Standardbedingungen mehr oder weniger handwerksmäßig durchführbar ist,
- drittens eine relative Positionsbestimmung des untersuchten Individuums innerhalb einer Gruppe oder in Bezug auf ein bestimmtes Kriterium, z.B. einem Lehrziel ermöglicht und
- viertens bestimmte empirisch – also verhaltens- und erlebnisanalytisch, phänomenologisch und nicht etwa rein begrifflich – abgrenzbare Eigenschaften, Verhaltensdispositionen, Fähigkeiten, Fertigkeiten oder Kenntnisse prüft.“

Was lernen wir aus dieser Definition? Das erste angesprochene Merkmal, die wissenschaftliche Begründbarkeit, zielt darauf ab, dass ein individualdiagnostischer Test ein Personenmerkmal prüft, das sich nach wissenschaftlichen Kriterien beschreiben und begründen lässt. Anders ausgedrückt: Ein guter Test braucht eine gute theoretische wissenschaftliche Basis. Dieser Punkt wird im Laufe des Kapitels noch einmal unter dem Stichwort „Validität“ besprochen. Das zweite Merkmal legt fest, dass ein Test immer in vergleichbarer Art und Weise und unter vergleichbaren Bedingungen durchgeführt, ausgewertet und interpretiert werden muss. Darauf werden wir unter den Stichworten „Objektivität“ und „Reliabilität“ weiter eingehen. Das dritte Merkmal zielt auf den in Kapitel 1.2 besprochenen Umstand, dass ein Rohwert nur unter Verwendung einer Bezugsnorm inhaltlich interpretierbar ist.

Wozu testen?

Das vierte Merkmal schließlich schneidet ein Problem an, dass das Grundproblem der pädagogischen Individualdiagnostik und zugleich Anlass für die Nutzung von Testverfahren ist: Die Personenmerkmale, deren Ausprägungen man kennen möchte, um darauf aufbauend pädagogische Entscheidungen treffen zu können, sind einer direkten Empirie nicht zugänglich, d.h. sie sind nicht direkt beobachtbar (Abbildung 1). Die mathematische Fähigkeit eines Schülers oder auch seine Intelligenz ist dem Schüler nicht direkt anzusehen. Ob ein Schüler motiviert ist oder ängstlich ist, kann man nicht direkt beobachten. Was man beobachten kann sind bspw. niedergeschriebene Antworten eines Schülers in einem Mathematikleistungstest. Man kann auszählen, wie viele richtige Antworten ein Schüler in einem Intelligenztest angekreuzt hat. Man kann das mehr oder weniger ängstliche Verhalten eines Schülers beobachten oder man kann aus der Tatsache, dass ein Schüler innerhalb eines Projektes ein hohes Engagement zeigt, auf eine entsprechend hohe Motivation schließen. Das bedeutet: Wenn wir einen Test einsetzen, dann wollen wir eine Information über ein Merkmal eines Schülers, das wir nicht sehen können. Man spricht in dem Fall von einem „latenten“ Merkmal. Der Test liefert uns Information über das Schülermerkmal, in dem er uns unter möglichst standardisierten Bedingungen ein bestimmtes Verhalten (bspw. die niedergeschriebene Lösung einer Testaufgabe) eines Schülers beobachten lässt. Dieses direkt beobachtbare Verhalten bezeichnet man als „manifestes“ Merkmal. Anhand dieses Verhaltens

Latentes und manifestes Merkmal

schließen wir dann auf die Ausprägung des nicht direkt beobachtbaren, latenten Schülermerkmals. Für diesen Schluss – von dem beobachteten Verhalten auf die Ausprägung des nicht direkt beobachtbaren Personenmerkmals – ist jedoch eine Annahme notwendig, nämlich dass das beobachtete Verhalten maßgeblich von dem interessierenden latenten Personenmerkmal beeinflusst ist. Inwiefern diese notwendige Annahme berechtigt ist und unter welchen Bedingungen man davon ausgehen kann, dass diese Annahme in einer bestimmten Testsituation gültig ist, ist Gegenstand von Testtheorien, also einem Bündel theoretischer Annahmen über das Zusammenspiel nicht beobachtbarer Personenmerkmale und beobachtbarem Verhalten in Testsituationen. Die bekannteste Testtheorie, die sogenannte „Klassische Testtheorie“ wird im Folgenden weiter ausgeführt. Eine weitere Testtheorie, die insbesondere im Rahmen sogenannter „large-scale assessments“ Anwendung findet, können Sie ausführlich in den Studienbriefen „Vergleichsarbeiten“ und „Bildungsmonitoring“ kennen lernen.

Testtheorie

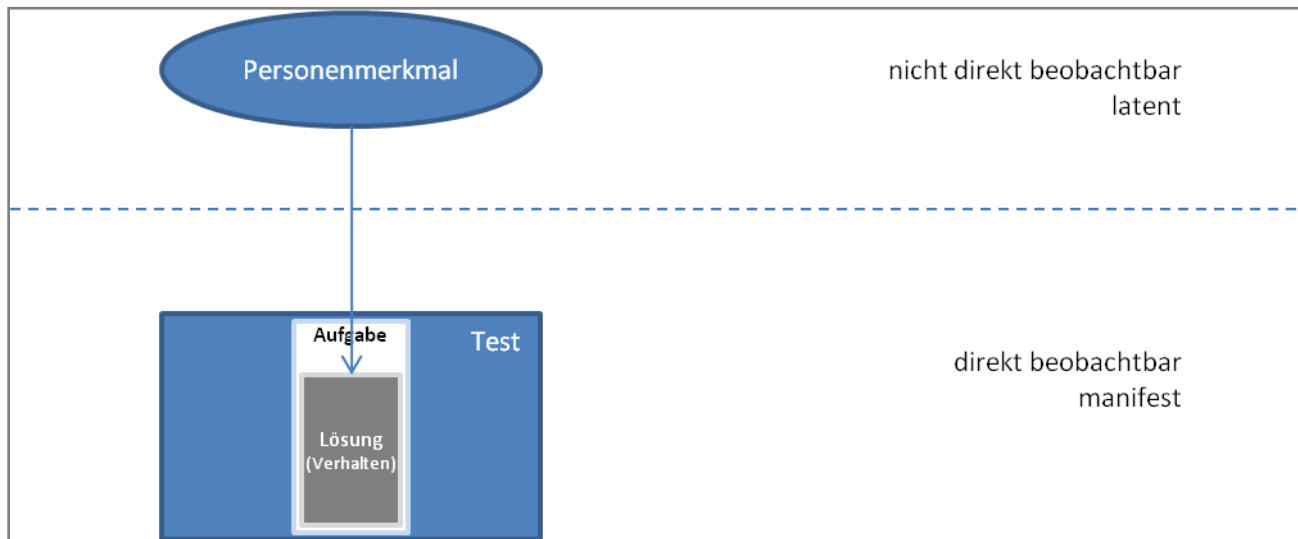


Abbildung 1: Annahme über den Zusammenhang zwischen dem Personenmerkmal und dem Lösen einer Testaufgabe

1.3.2 Das Testergebnis

Angenommen, wir wollten überprüfen, ob ein Schüler dazu in der Lage ist, einen für seine Altersgruppe angemessenen Text verstehend zu lesen. Dafür würden wir ihm einen entsprechenden Text geben, mit der Bitte ihn innerhalb einer bestimmten Zeit zu bearbeiten und zu versuchen, den Textinhalt zu verstehen. Um das Textverständnis zu überprüfen, würden wir dem Schüler nach dem Lesen drei Fragen zum Text stellen, die er kurz schriftlich beantworten soll (Abbildung 2). Das zu überprüfende latente Personenmerkmal wäre in diesem Beispiel das Textverständnis (im Sinne der Fähigkeit, einen Text verstehend zu lesen). Der Test bestünde aus dem Text und den drei Fragen. Das beobachtbare manifeste Verhalten wären die drei Antworten auf die drei Fragen.

Unsere grundlegende Annahme wäre, dass das Textverständnis des Schülers dafür verantwortlich ist, ob wir von ihm gute und richtige Antworten geliefert bekommen oder nicht. Diese sehr einfache, aus genau einer Annahme bestehende Testtheorie ist jedoch leider nicht haltbar. Die Annahme, dass die Antwort in einem Test ausschließlich von der Fähigkeit einer Person abhängt, ignoriert, dass es noch viele weitere Einflüsse auf das in einem Test gezeigte Verhalten geben kann. Bspw. könnte der Schüler zwar über ein hohes Textverständnis verfügen, jedoch wenig motiviert sein, sich testen zu lassen, weshalb seine Antworten nicht so gut ausfallen wie möglich. Oder aber der Schüler könnte bei einer Frage zwar unsicher gewesen sein, aber trotzdem mit etwas Glück eine gute Antwort aufgeschrieben haben. Evtl. war der Schüler beim Lesen eines Textabschnitts durch einen Mitschüler abgelenkt, so dass er, bei eigentlich gutem Textverständnis, für eine diesen Abschnitt betreffende Frage keine gute Antwort abliefern konnte. Die Anzahl verschiedener möglicher Einflüsse auf das in einem Test gezeigte Verhalten ist nahezu unendlich groß. Aus diesem Grund ist es auch müßig zu versuchen, jeden einzelnen Einfluss genau zu benennen, zumal die Hoffnung besteht, dass jeder einzelne dieser Einflüsse verschwindend gering ist. Deshalb werden diese Einflüsse üblicherweise unter dem Begriff „Fehler“ einfach zusammengefasst. Welche Eigenschaften dieser Fehler hat, wie er das Testverhalten beeinflusst, wie man seine Größe bestimmen kann und wie man Tests konstruieren kann, bei denen der Einfluss des Fehlers auf das Testverhalten möglichst gering ist, das ist Gegenstand der sogenannten „Klassischen Testtheorie“.

Die Klassische Testtheorie besteht aus einem Satz von Aussagen (Axiome), der den theoretischen Hintergrund für die meisten individualdiagnostischen Verfahren bildet. Kennzeichnend für die Klassische Testtheorie ist zum einen die Annahme, dass das Testergebnis zwar maßgeblich von der Ausprägung des latenten Personenmerkmals abhängt, dass es aber zusätzlich noch einen „Messfehler“ beinhaltet. Da dieses eine der zwei zentralen Annahmen der Klassischen Testtheorie ist, wird sie oftmals auch als Fehlertheorie bezeichnet.

Zufällige
Einflüsse =
Fehler

Messwert =
wahrer Wert +
Messfehler

Der Messfehler (kurz: Fehler) repräsentiert den zufälligen und unsystematischen Anteil des Testergebnisses, der nicht auf das latente Personenmerkmal, sondern auf situative Zufälligkeiten zurückgeführt werden muss (Gröschke, 2005). Gemäß der Klassischen Testtheorie setzt sich damit ein Messwert, sprich das Testergebnis, immer aus zwei Anteilen additiv zusammen: dem wahren Wert, der auf die Ausprägung des Personenmerkmals zurückgeführt werden kann, und dem Messfehler.

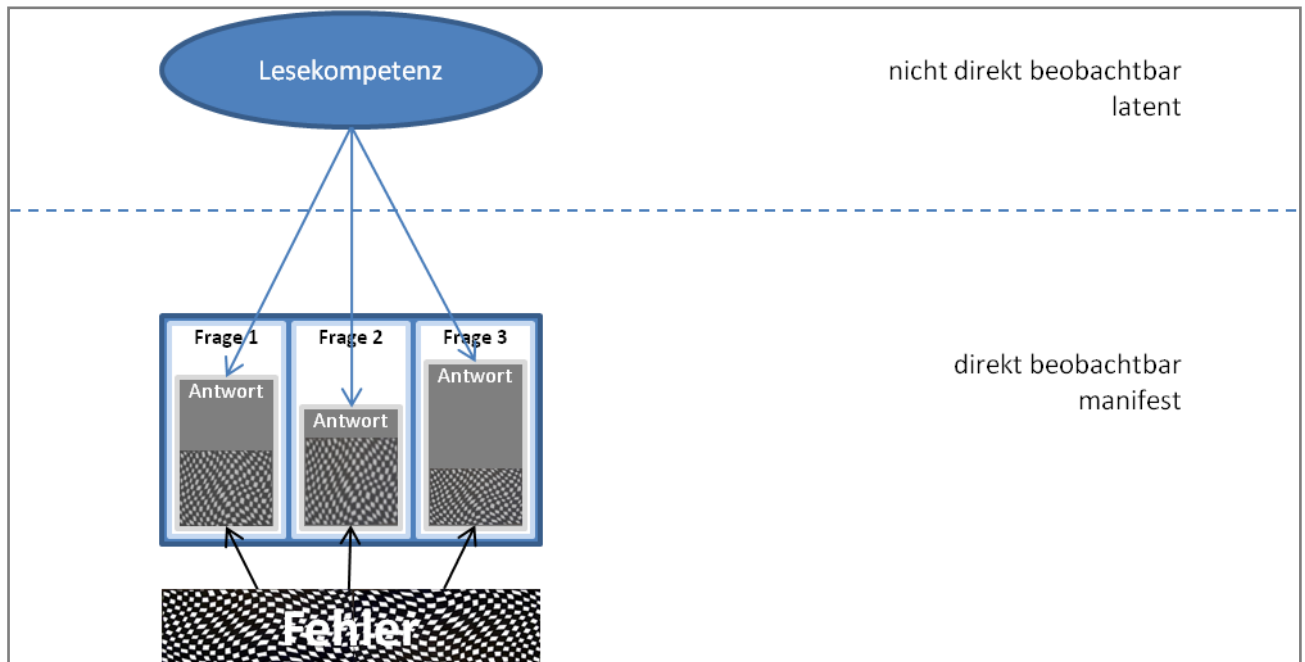


Abbildung 2: Einfluss des Personenmerkmals sowie zufälliger Gegebenheiten (Fehler) auf das Verhalten in Tests

Fehler = Zufall

Die zweite zentrale Annahme der Klassischen Testtheorie ist, dass der Messfehler, wie bereits erwähnt, rein zufällig ist. Aus dieser Annahme ergibt sich, dass der Messfehler unabhängig ist von jeglichen Gegebenheiten. Wie hoch der Anteil eines Messergebnisses ist, das auf zufällige Umstände zurückgeführt werden muss, ist bspw. unabhängig von der Ausprägung des eigentlich interessierenden Personenmerkmals, dem latenten Merkmal. Er ist unabhängig von der Ausprägung weiterer Merkmale der Person, und er ist unabhängig von Messfehlern, die bei früheren Einsätzen des Testinstruments zu verzeichnen waren. Diese (durch die Klassische Testtheorie postulierte) Eigenschaft des Fehlers ist von zentraler Bedeutung. Aus ihr lässt sich herleiten, wie der Fehleranteil an einem Testergebnis reduziert werden kann. Dieses werden wir im nächsten Kapitel „Indikatorbildung“ besprechen. Zum anderen macht man sich die Zufälligkeit des Fehlers zu Nutze, wenn für ein individualdiagnostisches Testverfahren überprüft werden soll, wie groß der Messfehleranteil am Testergebnis ist. Wir werden im Kapitel „Reliabilität“ auf diesen Punkt zurück kommen.

1.3.3 Indikatorbildung

Üblicherweise besteht ein Test nicht nur aus einer Aufgabe, sondern aus mehreren bis vielen Testaufgaben. Bspw. besteht unser Leseverständnistest nicht nur aus einer, sondern aus drei Aufgaben. Für jede Aufgabe können wir beobachten, ob der Schüler sie korrekt bearbeitet oder nicht. Wurde eine Aufgabe gelöst, bekommt der Schüler dafür einen oder mehrere Punkte (=Messwerte), ansonsten eben nicht. Damit bekommen wir für jeden Schüler so viele Zahlen (Messwerte) wie der Test Aufgaben hat. Da wir die Ausprägung des Personenmerkmals – in unserem Fall das Ausmaß der Fähigkeit, einen Text verstehend zu lesen – aber durch nur eine Zahl ausdrücken möchten, berechnen wir üblicherweise die Summe oder den Mittelwert der Punkte über alle Testaufgaben hinweg. Diese Summe bzw. dieser Mittelwert ist dann das Testergebnis. Es dient als Indikator für die Ausprägung des latenten Personenmerkmals. Die Aufgaben, deren Punkte durch eine Summe oder einen Mittelwert zusammengefasst werden, bilden eine sogenannte „Skala“. Eine Skala ist die Messlatte, mit deren Hilfe ein Messwert bestimmt wird. Sie erstreckt sich von dem kleinsten Wert, den die Summe oder der Mittelwert annehmen kann, bis hin zu dem entsprechenden größten Wert. Bekäme bspw. in unserem Leseverständnistest ein Schüler bei jeder Aufgabe für ihre korrekte Bearbeitung genau einen Punkt (und ansonsten null Punkte), dann erhielten wir bei drei Aufgaben durch Summenbildung eine Skala mit Werten von 0 bis 3.

Skala

Reduktion des
Fehlers

Die Summenbildung hat jedoch nicht nur das Ziel, mit nur möglichst einem Indikator die Ausprägung des Personenmerkmals einschätzen zu können. Viel wichtiger ist, dass auf diese Art der Fehleranteil am Testergebnis reduziert wird. Und der Grund dafür liegt in der Zufälligkeit des Fehlers: Die Antwort jeder Aufgabe, sprich jeder Messwert, enthält einen Fehler. Dabei ist es aber zum einen vollkommen zufällig, wie groß dieser Fehler ist. Zum anderen ist es

ebenfalls vollkommen zufällig, ob der Fehler dazu führt, dass der Messwert den wahren Wert überschätzt oder unterschätzt (ob also der Fehler zum wahren Wert addiert oder von ihm subtrahiert werden muss). Bildet man jetzt die Summe über mehrere Messwerte, die jeweils entweder einen „überschätzenden“ oder einen „unterschätzenden“ Fehler enthalten, dann subtrahiert man automatisch von der Summe der überschätzenden Fehler die Summe der unterschätzenden Fehler. Dadurch reduziert sich automatisch der Anteil der Fehler an der Summe der Messwerte, sprich am Testergebnis. Im Idealfall ist die Summe der überschätzenden Fehler gleich der Summe der unterschätzenden Fehler. In dem Fall würden sich beide Fehlersummen gegenseitig aufheben und der Anteil des Fehlers am Testergebnis wäre Null.

Dieser Idealfall ist natürlich mehr als selten. Aber man kann nahe an ihn herankommen. Der Trick ist, die Anzahl der Testaufgaben zu erhöhen. Je mehr Aufgaben ein Test enthält, desto höher ist die Wahrscheinlichkeit, dass die Summe der überschätzenden Fehler gleich der Summe der unterschätzenden Fehler wird. Anders ausgedrückt, je mehr Aufgaben ein Test enthält, desto höher ist die Wahrscheinlichkeit, für einen Messwert mit einem Fehleranteil von $+x$ einen Messwert mit einem Fehleranteil von $-x$ zu finden. Summiert man beide Fehleranteile, so ergibt sich Null.

Viele
Aufgaben =
wenig Fehler

1.3.4 Kriterien der Qualität von Tests

Nachdem wir uns erarbeitet haben, was ein individualdiagnostischer Test ist und welche Eigenschaften ein Ergebnis hat, wenden wir uns nun der Frage zu, woran man die Qualität eines Tests erkennen kann bzw. auf welche Art und Weise man diese gewährleisten und überprüfen kann. Bei der Bewertung der Qualität von Tests orientiert man sich zunächst einmal hauptsächlich an drei Testgütekriterien, nämlich an der Reliabilität, der Validität sowie der Objektivität. Wir werden diese im Folgenden genauer besprechen. Neben diesen sogenannten „Hauptgütekriterien“ gibt es jedoch auch eine Menge weiterer sogenannter „Nebengütekriterien“. Zu den wichtigsten zählt dabei sicherlich die Normierung eines Tests. Normierungen von Tests werden Gegenstand des nachfolgenden Kapitels 1.3.4 sein.

1.3.4.1 Reliabilität

Wie bereits erwähnt, nimmt die Klassische Testtheorie an, dass ein Testergebnis zwar ein guter Indikator für die latente Merkmalsausprägung sein kann, dass das Ergebnis jedoch durch den Messfehler verzerrt wird. Wie hoch der Fehleranteil am Testergebnis ist, das ist die Frage nach der sogenannten Reliabilität des Tests. Die Reliabilität eines Tests spiegelt die Zuverlässigkeit oder die Genauigkeit eines Tests wider. Sie ist umso höher, je geringer der Fehleranteil am Testergebnis ist.

Definition: Reliabilität

Die Reliabilität eines Tests kennzeichnet den Grad der Genauigkeit, mit dem das geprüfte Merkmal gemessen wird (Bortz, 2005).

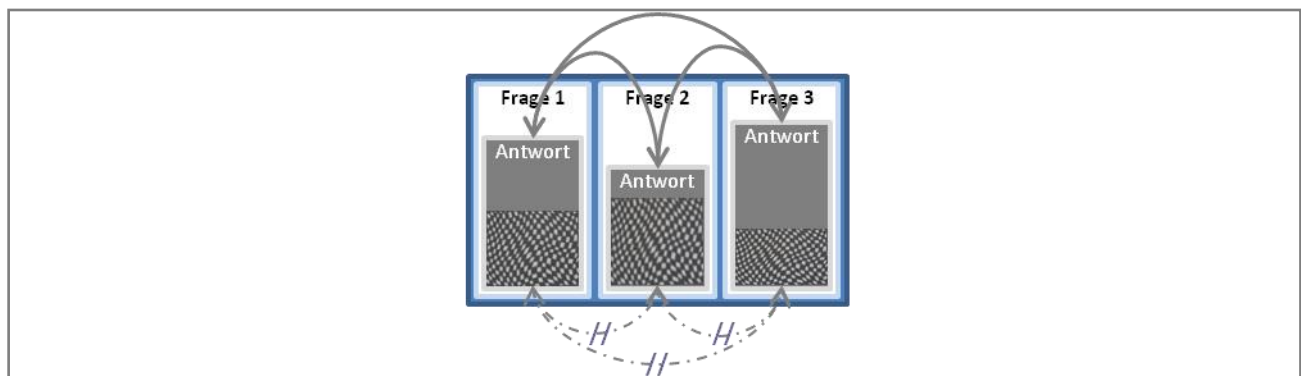


Abbildung 3: Reliabilität - Zusammenhänge der wahren Werte und Unabhängigkeiten der Messfehler

Es stellt sich jedoch die Frage, wie die Reliabilität, d.h. die Messgenauigkeit eines Tests ermittelt werden kann. Die Annahmen der Klassischen Testtheorie bieten die Grundlage für eine Antwort. Eine der beiden zentralen Annahmen der Klassischen Testtheorie ist, dass die Ausprägung des Messfehlers rein zufällig ist. Das bedeutet, dass der Messfehler von allen Gegebenheiten unabhängig ist, sprich der Messfehler steht in keinerlei Zusammenhang mit irgendetwas anderem (Kap. 1.3.2). Anders verhält es sich jedoch mit dem Anteil am Messwert, der durch die Ausprägung des zu messenden latenten Personenmerkmals bedingt ist, also dem wahren Wert. Der wahre Wert ist abhängig von der Ausprägung des Personenmerkmals, sprich der wahre Wert steht in systematischem Zusammenhang mit dem Personenmerkmal. Dadurch steht dieser wahre Wert jedoch auch in systematischem Zusammenhang mit den wahren Werten von Messergebnissen, die beim Testen desselben Personenmerkmals mit verschiedenen Aufgaben oder aber bei einer wiederholten Durchführung desselben Tests erzielt wurden.

Die wahren Werte stehen in starkem Zusammenhang (sprich: sie sind korreliert), die Messfehler stehen in keinem Zusammenhang (Abbildung 3). Diese zentrale Annahme der Klassischen Testtheorie kann man sich zu Nutze machen, um die Reliabilität eines Testverfahrens zu bestimmen. Das Grundprinzip ist dabei immer dasselbe: Es wird überprüft, wie stark der Zusammenhang zwischen wiederholten oder verschiedenen Messungen desselben Personenmerkmals ist. Je höher dieser Zusammenhang ist, desto höher muss der Anteil der wahren Werte an den Messergebnissen sein, da ausschließlich die wahren Werte in Zusammenhang zueinander stehen. Ein hoher Fehleranteil an den Messwerten würde dazu führen, dass die Messwerte in keinem Zusammenhang miteinander stehen. Die Messwerte wären unkorreliert.

Um das Ausmaß an Genauigkeit eines Tests ausdrücken zu können, wird meist ein sogenannter Korrelationskoeffizient berechnet. Dieser drückt durch eine Zahl die Stärke des Zusammenhangs zweier Variablen aus. Bspw. kann ein Korrelationskoeffizient berechnet werden für den Zusammenhang zwischen den Punkten, die in zwei Testantworten erzielt wurden. Wurde derselbe Test mit denselben Personen zu zwei Zeitpunkten wiederholt durchgeführt, kann der Zusammenhang zwischen den Testleistungen zu den beiden Testzeitpunkten durch die Berechnung eines Korrelationskoeffizienten ausgedrückt werden.

Definition: Korrelationskoeffizient

Der Zusammenhang zwischen zwei Testergebnissen wird als Korrelation bezeichnet. Die Stärke des Zusammenhangs wird durch den Korrelationskoeffizienten ausgedrückt, der wiederum durch den Buchstaben r abgekürzt wird (z.B. $r = 0,65$). Dieser Korrelationskoeffizient r kann einen Wert zwischen -1 und $+1$ annehmen. Werte, die gegen $+1$ oder auch gegen -1 streben, zeigen einen starken Zusammenhang an, während Werte nahe Null für einen schwachen bis gar keinen Zusammenhang stehen. Das Vorzeichen enthält die Information über die Art des Zusammenhangs: Ein positives Vorzeichen steht für einen gleich gerichteten Zusammenhang (Je größer..., desto größer...), ein negatives Vorzeichen steht für einen entgegengesetzt gerichteten Zusammenhang (Je größer..., desto kleiner...).

Es gibt verschiedene Formen, die Reliabilität eines Tests zu bestimmen. Sie werden auf den folgenden Seiten kurz dargestellt. Das Grundprinzip ist dabei aber immer dasselbe. Über die Berechnung eines Korrelationskoeffizienten (oder aber einer bestimmten Abwandlung des Korrelationskoeffizienten) wird die Genauigkeit des Tests durch eine Zahl ausgedrückt und so bewertbar gemacht.

Test-Retest-Reliabilität (Stabilität)

Für die Bestimmung der Test-Retest-Reliabilität wird derselbe Test mit einer Gruppe von Personen in einem größeren zeitlichen Abstand zweimal durchgeführt (Abbildung 4). Man spricht daher auch von der „Testwiederholungsmethode“. Berechnet wird dann der Korrelationskoeffizient (kurz: die Korrelation) für den Zusammenhang zwischen den Ergebnissen beider Testzeitpunkte. Dabei entsteht ein hoher Zusammenhang, wenn die Personen, die beim ersten Testzeitpunkt ein hohes Ergebnis erzielen konnten, dieses auch beim zweiten Testzeitpunkt erzielen, und wenn in gleichem Maße Personen, die beim ersten Mal eher niedrige Ergebnisse erreichten, auch beim zweiten Mal eher niedrige Testwerte erreichen.

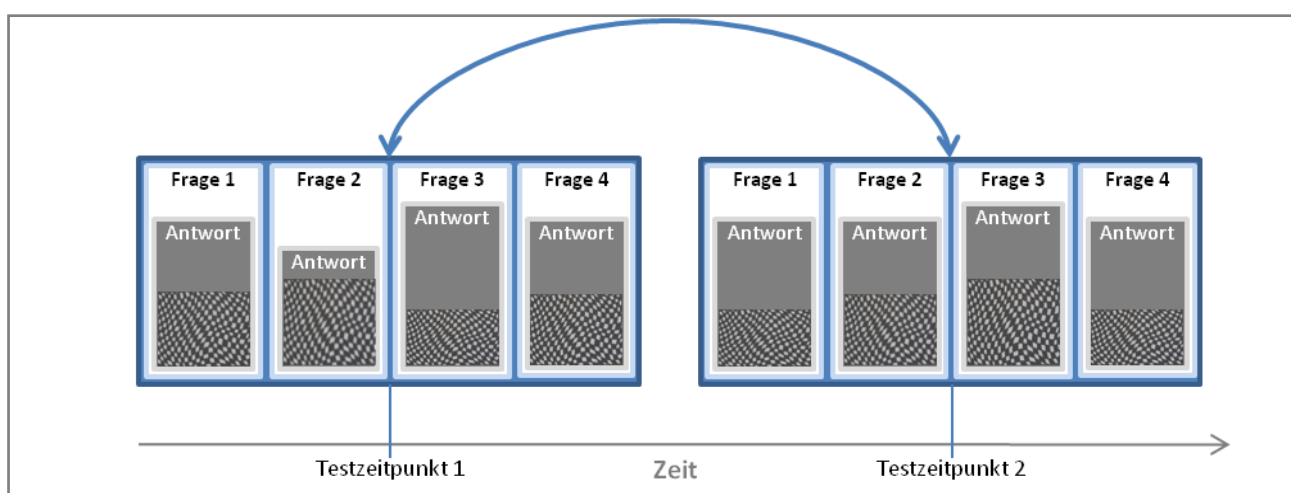


Abbildung 4: Test-Retest-Reliabilität

Beispiel: Test-Retest-Reliabilität

Angenommen, unserer Lehrer aus den Beispielen in Kapitel 1.2 hätte Zweifel an der Reliabilität seines Lesegeschwindigkeitstests. Um diese zu überprüfen, würde er den Test in einer seiner Schulklassen in einem zeitlichen Abstand von drei Wochen zweimal durchführen. Für jeden Testzeitpunkt könnte er dann zunächst seine Schüler nach der erreichten Punktzahl im Test sortieren, sie also in eine Rangreihe bringen. Wenn die Rangreihen für die beiden Testzeitpunkte sehr stark korrespondierten, dann spräche das für eine gute Reliabilität des eingesetzten Lesegeschwindigkeitstests. Statt dem Vergleichen von Rangreihen könnte der Lehrer aber natürlich auch die Korrelation berechnen (Wie das mit Hilfe von Microsoft Excel recht einfach geht, ist Gegenstand von Kapitel 1.5). Wenn die beiden Rangreihen sehr ähnlich oder identisch sind, dann strebt der Korrelationskoeffizient gegen $r = +1$.

Aus dem Beispiel werden wenigstens zwei Bedingungen deutlich, die gegeben sein müssen, um die Testhalbierungsmethode einsetzen zu können. Zum einen sollte man von dem Personenmerkmal, das gemessen werden soll – im Beispiel also die Lesegeschwindigkeit –, annehmen können, dass es sich nicht über die drei Wochen hinweg verändert. Zum anderen muss die Zeitspanne zwischen den beiden Testzeitpunkten groß genug sein, um ausschließen zu können, dass das Durchführen des ersten Tests Einfluss auf die Ergebnisse des zweiten Tests hat. Dies könnte bspw. durch Lerneffekte der Fall sein.

Paralleltest-Reliabilität (Äquivalenz)

Die Methode kommt dann zum Einsatz, wenn aus praktischen Gründen zwei äquivalente Testversionen erforderlich sind, bspw. um zu verhindern, dass die Testpersonen von ihren Nachbarn die Lösungen abschreiben (Abbildung 5). Um die Paralleltest-Reliabilität zu ermitteln, werden zwei Versionen eines Tests entwickelt, die beide gleichermaßen auf die Erfassung desselben Personenmerkmals abzielen. Beide Tests werden mit denselben Testpersonen in kurzem zeitlichem Abstand innerhalb einer Sitzung durchgeführt. Im Anschluss wird die Korrelation der beiden Testversionen berechnet. Diese Methode ist bei der Entwicklung eines Tests natürlich recht aufwändig, da man streng genommen nicht nur einen, sondern zwei Tests entwickeln muss. Sie bietet sich entsprechend nur an, wenn man aus praktischen Gründen (s.o.) sowieso zwei Testversionen braucht.

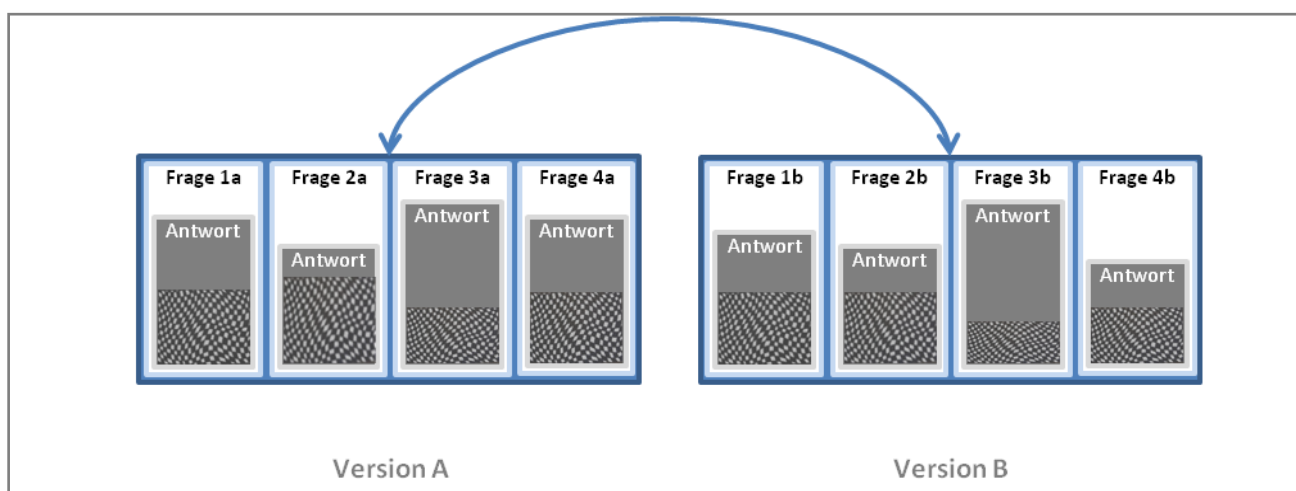


Abbildung 5: Paralleltest-Reliabilität

Testhalbierungs-Reliabilität

Im Gegensatz zur Paralleltest-Methode ist die Testhalbierungsmethode mit geringem Zusatzaufwand für die Testentwicklung verbunden. Der Test wird dafür nur einmal mit einer Gruppe von Personen durchgeführt. Im Anschluss werden die Aufgaben des Tests in zwei Gruppen aufgeteilt (z.B. „split-half-Methode“: die erste Hälfte der Aufgaben versus die zweite Hälfte der Aufgaben oder „odds-even-Methode“: die Aufgaben mit einer geraden Aufgabennummer versus die Aufgaben mit einer ungeraden Aufgabennummer). Die beiden Aufgabengruppen werden dann wie zwei separate Tests behandelt, und es wird für jede Aufgabengruppe das Testergebnis berechnet. Danach berechnet man, ähnlich wie bei der Paralleltest-Methode, die Korrelation dieser beiden Testergebnisse (Abbildung 6).

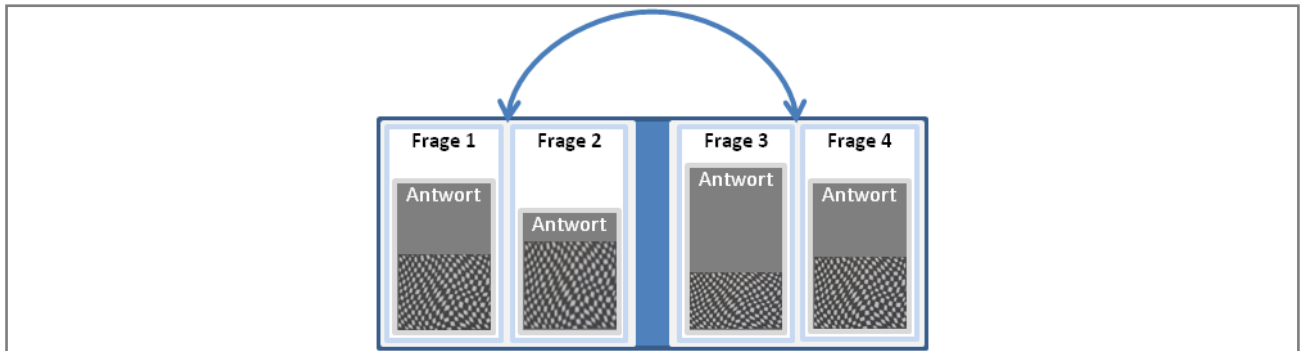


Abbildung 6: Testhalbierungs-Reliabilität

Interne Konsistenz

Die sogenannte „interne Konsistenz“ ist ebenfalls ein Schätzer für die Reliabilität eines Tests. Der Gedanke dahinter ist, dass die Antworten einer Person auf Aufgaben, die alle dasselbe Personenmerkmal abbilden sollen, von ähnlicher Qualität sein sollten. Entsprechend sollten alle Aufgaben eines Tests in engen Zusammenhängen zueinander stehen, was als interne Konsistenz des Tests bezeichnet wird. Bestimmt wird die interne Konsistenz, indem die Korrelationen einer jeden Aufgabe mit allen anderen Aufgaben berechnet (Abbildung 7) und unter Berücksichtigung bestimmter Gewichte zu einem Mittelwert zusammengefasst werden. Der resultierende Koeffizient nennt sich „Cronbachs Alpha“ (Cronbach, 1951). Auch er kann theoretisch Werte zwischen -1 und +1 annehmen. Negative Werte sind jedoch äußerst selten. Von einer hohen internen Konsistenz geht man aus, wenn Cronbachs $\alpha \geq +0,8$ ist.

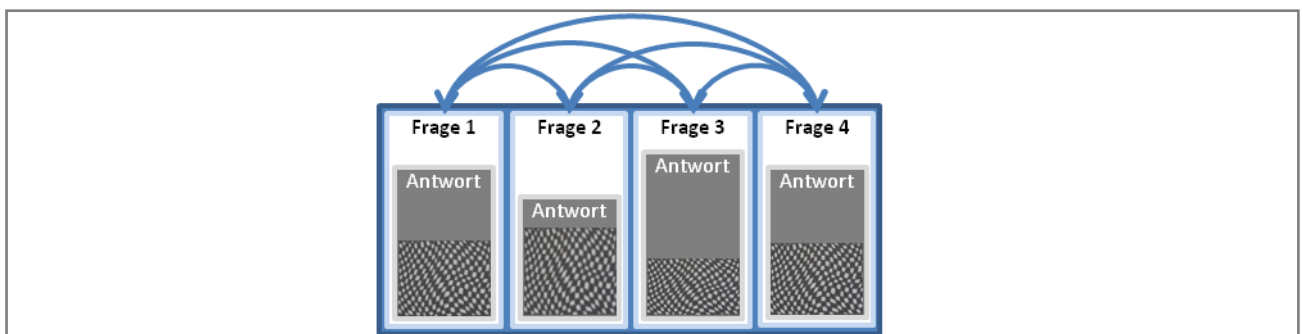


Abbildung 7: Interne Konsistenz

1.3.4.2 Validität

Die Validität eines Tests ist das Merkmal, das in der Lienertschen Definition eines Tests (Kap. 1.3.1) unter dem Stichwort „wissenschaftliche Begründbarkeit“ als erstes angesprochen wurde, was die Bedeutung dieses Testgütekriteriums unterstreicht. Sie betrifft die Frage, ob die Antworten in einem Test tatsächlich maßgeblich von dem vermuteten Personenmerkmal abhängen oder aber von einem ganz anderen. Sind die Antworten im Leseverständnistest tatsächlich von der Lesekompetenz der Schülerinnen und Schüler geprägt oder gibt es zusätzlich oder alternativ dazu weitere Personenmerkmale, die die Antworten im Lesekompetenztest systematisch beeinflussen? Welchen Einfluss hat bspw. die Motivation eines Schülers, in diesem Test eine möglichst gute Leistung zu zeigen (Abbildung 8)? Oder nehmen wir einen Test zur Erfassung mathematischer Kenntnisse und Fähigkeiten, der Textaufgaben enthält. Sind die Antworten auf diese Aufgaben wirklich ausschließlich durch die mathematischen Kenntnisse und Fähigkeiten des Schülers bedingt oder werden sie zusätzlich von dessen Lesekompetenz beeinflusst? Und falls ja, wie stark ist dann dieser zusätzliche Einfluss der Lesekompetenz? Ist er so gering, dass man ihn vernachlässigen kann oder so stark, dass man nicht mehr davon ausgehen kann, dass die manifesten Testantworten gute Indikatoren für die latenten mathematischen Fähigkeiten sind?

Definition: Validität

Die Validität (Gültigkeit) eines Tests gibt an, wie gut der Test in der Lage ist, genau das zu messen, was er zu messen vorgibt (Bortz, 2005).

Die grundlegende Voraussetzung für die Validität eines Tests ist die klare theoretische Definition des Personenmerkmals, das getestet werden soll. In Bezug auf den Lesekompetenztest z.B. müssen wir uns fragen, was genau wir unter Lesekompetenz verstehen? Ist Lesekompetenz gleich Leseverständnis? Und was genau ist Leseverständnis?

Gehört zur Lesekompetenz neben dem Leseverständnis auch die Lesegeschwindigkeit? Sich solche und andere Fragen zu stellen ist insbesondere für zwei Punkte wichtig. Zum einen ist die Beantwortung dieser Fragen eine notwendige Voraussetzung, wenn ein individualdiagnostisches Testverfahren neu entwickelt werden soll. Inwieweit im schulischen Kontext dies nötig ist und wie in einem solchen Fall diese Fragen beantwortet werden können, ist Gegenstand des Kapitels 1.5. Zum anderen müssen diese Fragen beantwortet werden, wenn aus der Fülle bereits existierender Testverfahren eines ausgesucht werden soll, das einem genau die Informationen liefert, die für eine pädagogische Entscheidung benötigt werden (Für die Suche und Auswahl individualdiagnostischer Testverfahren im pädagogischen Kontext siehe die UDiKom-Testdatenbank, die im Internet unter <http://tests.udikom.de/> frei verfügbar ist).

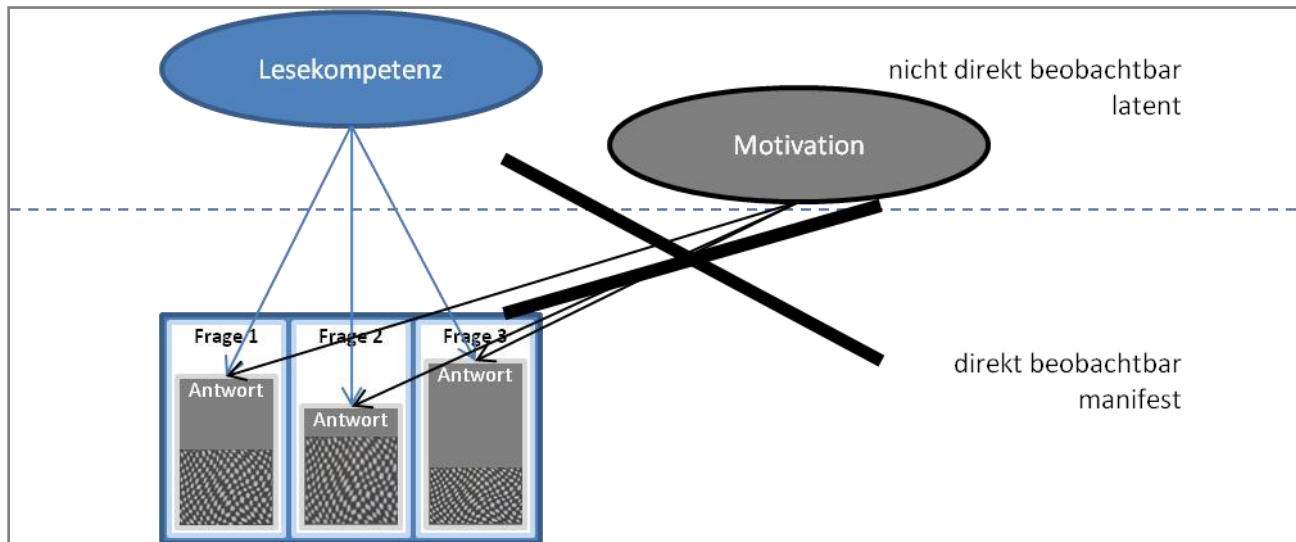


Abbildung 8: Validität

Doch wie kann man diese Fragen beantworten? Hier kommt die wissenschaftliche Begründbarkeit ins Spiel. Die entsprechenden wissenschaftlichen Disziplinen wie die Erziehungswissenschaft oder die Pädagogische Psychologie liefern Theorien und Modelle, mit deren Hilfe die latenten Personenmerkmale (vgl. Kap. 1.3.1) beschrieben werden können. Für eine gute pädagogische Individualdiagnostik ist es unerlässlich, sich vorab auf der Basis solcher wissenschaftlich begründeter theoretischer Modelle ein Bild von dem nicht direkt beobachtbaren Personenmerkmal zu machen. Dies ist die notwendige Grundlage, um entweder dazu passende Testverfahren auszuwählen oder aber selbst zu entwickeln.

Die unterschiedlichen Theorien und Modelle lassen sich anhand verschiedener Merkmale klassifizieren, wobei wir uns im Rahmen dieses Studienbriefs auf die Besprechung eines Merkmals beschränken wollen, nämlich das Merkmal der Dimensionalität. Die Dimensionalität betrifft die Frage, wie komplex ein Personenmerkmal ist bzw. wie komplex das beschreibende Modell sein muss. Ist bspw. die Lesekompetenz ein sehr wenig komplexes Personenmerkmal oder muss man innerhalb dieser Kompetenzen wiederum verschiedene Teilkompetenzen unterscheiden wie z.B. Leseverständnis und Lesegeschwindigkeit? Je mehr Teilkompetenzen oder Facetten durch ein Modell beschrieben werden, um das interessierende Personenmerkmal umfassend abbilden zu können, desto mehr Dimensionen hat das Modell. Für das entsprechende Testverfahren bedeutet dies, dass es nicht nur aus einer Skala besteht, sondern aus genau so vielen Skalen (vgl. Kap. 1.3.3) wie das Modell Dimensionen hat. Diese Skalen sind Untertests, deren Qualität genauso überprüft und bewertet werden muss, wie die entsprechende Qualität eines eindimensionalen Tests.

Unabhängig davon, ob ein Modell und damit ein Test eindimensional oder mehrdimensional ist, muss gewährleistet sein, dass der Test valide ist. Im Rahmen des Studienbriefs wollen wir drei gängige Arten der Validierung besprechen. Die erste ist die sogenannte „Inhaltsvalidität“. In diesem Fall geht es weniger um eine Überprüfung der Validität, sondern stärker um das Vorgehen bei der Entwicklung von Testverfahren, die die Validität gewährleisten sollen. Die weiteren beiden Validierungsarten, die „Kriteriumsvalidität“ sowie die „Konstruktvalidität“, sind dann jedoch Arten der empirischen Überprüfung der Validität. Hierfür müssen der zu validierende Test und andere Tests und Erfassungsmethoden bei einer Stichprobe von Personen eingesetzt werden, was einen gewissen Aufwand bedeutet.

Inhaltsvalidität

Die Inhaltsvalidität, auch „Kontentvalidität“ genannt, ist besonders bei Testverfahren relevant, die sich einem bestimmten Fach, einem Themenbereich oder auch einer bestimmten Unterrichtseinheit zuordnen lassen. Sie gibt an, wie gut die Aufgaben eines Tests den zu testenden Inhaltsbereich repräsentieren. Hierbei geht es insbesondere um die Frage, ob alle relevanten Aspekte eines Inhaltsbereichs durch die Aufgaben umfassend abgebildet werden. Die

Inhaltsvalidität wird meist durch Expertenbefragungen oder ähnliche Verfahren eingeschätzt oder aber durch die Art der Aufgabenkonstruktion und -auswahl sichergestellt (vgl. Kap. 1.5).

Definition: Inhaltsvalidität

„Inhaltsvalidität ist gegeben, wenn der Inhalt der Testitems das zu messende Konstrukt in seinen wichtigsten Aspekten erschöpfend erfasst [...]. Hieraus folgt jedoch, dass die Grundgesamtheit der Testitems, die potentiell für die Operationalisierung eines Items in Frage kommen, sehr genau definiert werden muss. Die Inhaltsvalidität eines Tests ist umso höher, je besser die Testitems diese Grundgesamtheit repräsentieren“ (Bortz & Döring, 2006, S. 200).

Kriteriumsvalidität

Bei der Kriteriumsvalidität wird überprüft, inwiefern eine Testleistung (und damit das interessierende latente Personenmerkmal) in Zusammenhang mit einem anderen direkt beobachtbaren manifesten Personenmerkmal steht. Es geht also, wie bereits bei der Bestimmung der Reliabilität, um die Stärke eines Zusammenhangs, die sich empirisch ermitteln und mittels eines Korrelationskoeffizienten numerisch angeben lässt. Im Gegensatz zur Reliabilitätsbestimmung wird jetzt jedoch nicht der Zusammenhang zwischen Aufgaben (bzw. Antworten) desselben oder wiederholt durchgeführten Tests überprüft, sondern der Zusammenhang zwischen dem Testergebnis und einem Personenmerkmal, das nicht durch den eigentlichen Test erfasst wird (Abbildung 9). Zudem handelt es sich bei diesem Personenmerkmal um ein manifestes, sprich direkt beobachtbares Merkmal.

Definition: Kriteriumsvalidität

Ein Test weist Kriteriumsvalidität auf, wenn vom Verhalten der Testperson innerhalb der Testsituation erfolgreich auf ein Kriterium, nämlich auf ein Verhalten außerhalb der Testsituation, geschlossen werden kann (Moosbrugger & Kevala, 2008).

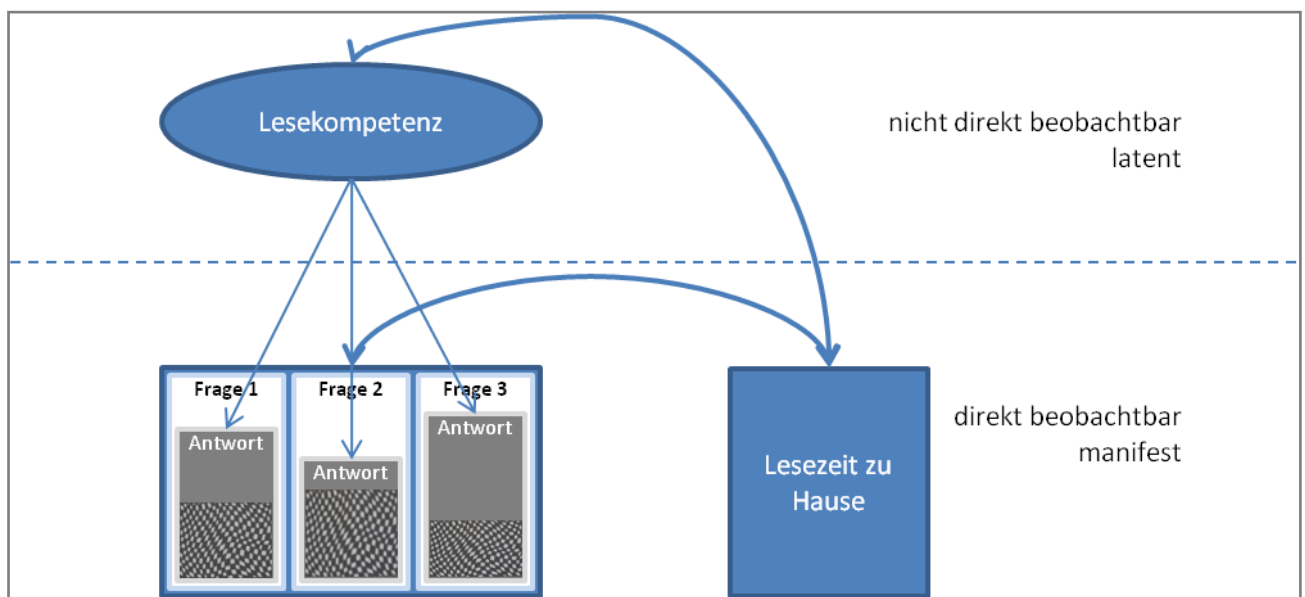


Abbildung 9: Kriteriumsvalidität

Beispiel: Kriteriumsvalidität

Es ist bekannt, dass Schülerinnen und Schüler, die zu Hause gerne und viel lesen, über eine höhere Lesekompetenz verfügen als Schülerinnen und Schüler, die nur ungern lesen. Wenn man die Kriteriumsvalidität eines Lesefähigkeitstests überprüfen wollte, könnte man entsprechend untersuchen, ob sich empirisch eine Korrelation zwischen dem Testergebnis und der durchschnittlichen Lesezeit zu Hause nachweisen lässt (Die Zeit, die jemand mit Lesen verbringt, ist direkt beobachtbar). Je höher der Korrelationskoeffizient, desto höher wäre die Kriteriumsvalidität einzuschätzen.

Konstruktvalidität

Anstelle eines direkt beobachtbaren Außenkriteriums, wie bei der Kriteriumsvalidität, werden zur Überprüfung der Konstruktvalidität latente, nicht direkt beobachtbare Personenmerkmale herangezogen (Abbildung 10). Ihre Bezeichnung erhält die Konstruktvalidität daher, dass latente, nicht direkt beobachtbare Personenmerkmale, die die zentrale Rolle bei der Konstruktvalidierung spielen, üblicherweise auch als „Konstrukt“ bezeichnet werden. Auf der

Basis entsprechender theoretischer Modelle sowie wissenschaftlicher Erkenntnisse über das Zusammenspiel der damit beschriebenen Personenmerkmale werden Hypothesen über Zusammenhänge zwischen dem zu erfassenden und anderen latenten Merkmalen formuliert und überprüft. Dabei können sowohl Vermutungen über starke Zusammenhänge als auch Vermutungen über Unabhängigkeiten (kein Zusammenhang) aus der Literatur abgeleitet und überprüft werden. Geht man davon aus, dass zwischen zwei latenten Personenmerkmalen ein Zusammenhang besteht, spricht man von *konvergenter Validität*. Ist kein Zusammenhang zu vermuten, wird von *diskriminanter Validität* gesprochen.

Definition: Konstruktvalidität

Ein Test ist konstruktvalid, wenn aus dem zu messenden Zielkonstrukt Hypothesen ableitbar sind, die anhand der Testwerte bestätigt werden können (Bortz & Döring, 2006, S. 201).

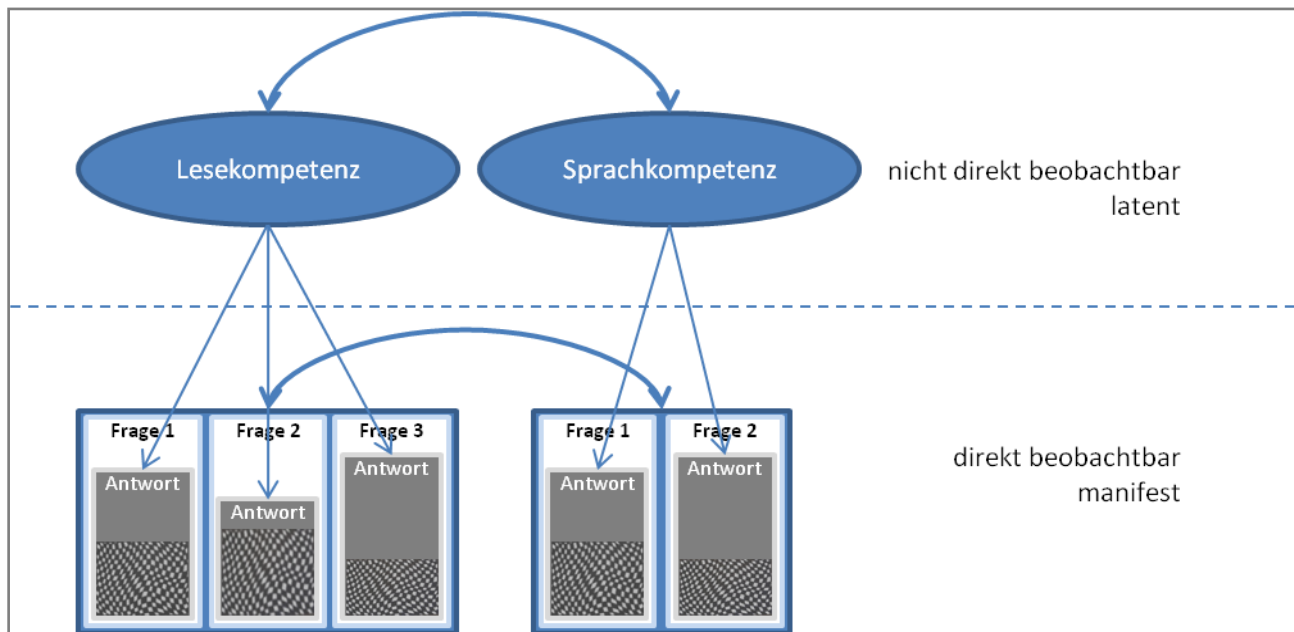


Abbildung 10: Konstruktvalidität

Beispiel: Konstruktvalidität

Aus der wissenschaftlichen Literatur ist bekannt, dass die Lesekompetenz von Schülerinnen und Schülern in engem Zusammenhang steht mit ihrer jeweiligen Sprachkompetenz. Die Annahme über den Zusammenhang zwischen den beiden latenten Personenmerkmalen „Lesekompetenz“ und „Sprachkompetenz“ kann im Rahmen einer konvergenten Validierung des Lesekompetenztests genutzt werden. Dafür setzen wir den Lesekompetenztest sowie einen Test zur Erfassung der Sprachkompetenz bei einer Stichprobe von Personen ein und berechnen hinterher die Korrelation zwischen den beiden Tests. Dieser empirisch beobachtete und durch den Korrelationskoeffizienten beschriebene Zusammenhang zwischen den beiden Tests sollte dem theoretisch angenommenen Zusammenhang zwischen den latenten Personenmerkmalen entsprechen. In unserem Beispiel würden wir also einen Korrelationskoeffizienten mit einem möglichst hohen positiven Betrag erwarten.

1.3.4.3 Objektivität

Die Objektivität eines Tests betrifft die Frage, in welchem Ausmaß ein Testergebnis abhängig ist vom Testanwender, also von der Person, die den Test durchführt (nicht zu verwechseln mit der Person, die getestet wird!). Es ist einleuchtend, dass die Antworten, die eine getestete Person in einem Test liefert, möglichst ausschließlich von dem betreffenden Merkmal der getesteten Person abhängen sollten, vollkommen unabhängig davon, wer mit ihr diesen Test durchführt. Bspw. sollten wir mit unserem Lesekompetenztest, den wir in einer Klasse einsetzen, bei jedem Schüler zu demselben Ergebnis gelangen, unabhängig davon, wer den Test in der Klasse durchführt, unabhängig davon, wer die Punkte für die niedergeschriebenen Antworten vergibt und unabhängig davon, wer aufgrund der Punktzahl entscheidet, ob einem Schüler eine hohe oder eine niedrige Lesekompetenz zugesprochen werden kann. Erreicht wird diese Unabhängigkeit vom Testanwender durch eine sogenannte *Standardisierung* des Tests. Sie ist gegeben, wenn in einer Testanleitung genau beschrieben wird, wie und unter welchen Bedingungen der Test durchgeführt werden muss, nach welchen Kriterien die Antworten im Test ausgewertet und mit Punkten versehen werden und wie die im gesamten Test erreichten Punkte (das Testergebnis) zu interpretieren sind (vgl. zum letzten Punkt Kap. 1.2). Ent-

sprechend unterscheidet man auch zwischen der *Durchführungsobjektivität*, der *Auswertungsobjektivität* sowie der *Interpretationsobjektivität*.

Durchführungsobjektivität

Die Durchführungsobjektivität ist gegeben, wenn ein Test immer auf dieselbe Art und Weise und unter denselben Bedingungen durchgeführt wird. Die Durchführungsobjektivität unseres Lesekompetenztests wäre bspw. gefährdet, wenn wir den Test einmal morgens während der ersten Schulstunde durchführten, ein anderes Mal jedoch in der letzten Nachmittagsstunde. Es wäre zu erwarten, dass die Testleistungen der Schüler in der ersten Stunde deutlich besser ausfallen als bei den Schülern, die am Ende eines langen Schultages erschöpft sind und entsprechend geringe Testleistungen zeigen. Die Bedingungen der Testdurchführung wären nicht miteinander vergleichbar und das Kriterium der Durchführungsobjektivität damit verletzt.

Auswertungsobjektivität

Die Auswertung der Testantworten, sprich die Entscheidung, wie viele Punkte der Schüler für eine Testantwort erhält, muss ebenfalls unabhängig sein von der Person, die diese Auswertung durchführt. Dabei gilt, je eindeutiger eine Antwort als richtig oder falsch, als gut oder schlecht bewertet werden kann, desto höher die Auswertungsobjektivität. Es gibt zwei Wege, diese Eindeutigkeit herzustellen. Zum einen können Aufgaben mit einem sogenannten „geschlossenen“ Antwortformat eingesetzt werden. Beispiele hierfür wären die bekannten Multiple-Choice-Aufgaben oder Lückentexte. Geschlossene Antwortformate zeichnen sich dadurch aus, dass vorab genau definiert werden kann, wie eine gute bzw. richtige Antwort aussieht. Im Falle von Multiple-Choice-Aufgaben ist genau definiert, welche der präsentierten Optionen angekreuzt werden muss (und welche nicht), um die volle Punktzahl zu erreichen. Bei Lückentexten ist vorab für jede Lücke genau ein Wort definiert, welches in die Lücke zu schreiben ist, um den entsprechenden Punkt zu erhalten.

Zum anderen können Auswertungsanleitungen und Schablonen erstellt werden, an die die auswertende Person sich möglichst genau hält. Genaue Auswertungsanleitungen sind insbesondere dann notwendig, wenn Aufgaben mit einem offenen Antwortformat eingesetzt werden. Offene Antwortformate zeichnen sich dadurch aus, dass es nicht genau eine richtige Antwort, sondern theoretisch unendlich viele richtige Antworten geben kann. Beispiel hierfür wären Fragen, die in Form eines mehr oder weniger langen Aufsatzes zu beantworten sind. In diesem Fall muss bei der Auswertung der Aufsatz nach fest vorgegebenen Kriterien (bspw. Argumentationsstruktur) und auf fest vorgegebene Art und Weise bewertet werden. Hierfür bieten sich z.B. Checklisten an, in denen die verschiedenen Kriterien als erfüllt oder nicht erfüllt abgehakt werden können.

Interpretationsobjektivität

Die Interpretationsobjektivität kann erhöht werden, indem Interpretationshilfen und -anweisungen zur Verfügung gestellt werden. Solche Hilfestellungen sind bspw. Normtabellen, wie wir sie in Kap. 1.2.2 unter dem Stichwort „Soziale Bezugsnorm“ besprochen haben. Sie stellen einen Vergleichsmaßstab zur Verfügung anhand dessen das Testergebnis einer Person in Bezug auf eine vergleichbare Gruppe von Personen bewertet und interpretiert werden kann. Interpretationshilfen in Bezug auf eine kriteriale Bezugsnorm könnten bspw. inhaltliche Beschreibungen verschiedener Abschnitte der Testskala sein. Auf diese Form wird im Studienbrief „Vergleichsarbeiten“ in Kapitel 2.4 detailliert eingegangen. Derartige Interpretationshilfen standardisieren die Testinterpretation und versuchen so, zu verhindern, dass individuelle Deutungen die Interpretation des Testergebnisses beeinflussen.

1.3.5 Nebengütekriterien

Die drei klassischen Testgütekriterien der Reliabilität, Validität und Objektivität werden durch sogenannte Nebengütekriterien ergänzt. Nebengütekriterien zeichnen sich durch einen starken Anwendungsbezug aus. Die Nebengütekriterien sind im Gegensatz zu den klassischen Gütekriterien uneinheitlich definiert. Die folgenden Kriterien werden in der Literatur häufig unter diesem Begriff zusammengefasst: Skalierung, Normierung, Testökonomie, Nützlichkeit, Unverfälschbarkeit und Fairness. Eine detaillierte Beschreibung dieser Kriterien findet sich bei Moosbrugger und Kevala (2008), die im Übrigen auf die begriffliche Abgrenzung zwischen Haupt- und Nebengütekriterien verzichten. Obwohl sämtliche der aufgeführten Nebengütekriterien relevant sind, würde deren Besprechung über den Rahmen dieses Kapitels hinausgehen. Daher soll in den folgenden Abschnitten einzig auf das Kriterium der Normierung eingegangen werden, welches für die Individualdiagnostik von zentraler Bedeutung ist.

1.3.5.1 Normierung

Die Ergebnisse eines Tests werden zunächst als Rohwert (Score) angegeben. Im Falle eines Leistungstests wäre dies bspw. die Anzahl der korrekt beantworteten Aufgaben. Um die Bedeutung dieser Rohwerte zu ermitteln, ist es erforderlich, die Ergebnisse in Relation zu den Ergebnissen einer vergleichbaren Referenzgruppe zu setzen (vgl. Kap. 1.2.2). Die Normierung eines Tests gibt auf Grundlage einer umfangreichen, repräsentativen Normierungsstichprobe Auskunft über die übliche Ausprägung des untersuchten Merkmals innerhalb einer Referenzgruppe. Die Testnormen können üblicherweise dem Testmanual bzw. den Handreichungen entnommen werden. Sollten bestimmte Faktoren wie Geschlecht und Alter für die Merkmalsausprägung relevant sein, so kann die Referenzgruppe anhand dieser Kriterien ausdifferenziert werden. In diesem Fall stehen entsprechende Geschlechts-, Alters- oder Schulnormen zur Verfügung. Die Normierungsdaten eines Tests bieten somit einen Bezugsrahmen zur Interpretation von Testergebnissen und ermöglichen durch den Vergleich mit einer Referenzgruppe eine Standortbestimmung einzelner Schüler.

Definition: Normierung

Das Nebengütekriterium der Normierung bewertet, inwieweit für die Ergebnisse eines Testinstruments Vergleichsdaten vorhanden sind, anhand derer sich Einzelergebnisse interpretieren lassen (Rost, 2004). Rost betont, dass ein normiertes Testinstrument nicht zwangsläufig auch den Hauptgütekriterien entspricht. Das Gütekriterium der Normierung ist unabhängig von der Objektivität, Reliabilität und Validität eines Instruments.

Normtabellen enthalten meist auch Angaben darüber, welchem Prozentrangplatz ein Testergebnis entspricht. Anhand des Prozentrangplatzes kann man ablesen, wie viele der Personen in der Normierungsstichprobe ein besseres und wie viele ein schlechteres Testergebnis erzielt haben. Bspw. bedeutet ein Prozentrangplatz 87, dass 13% der Personen in der Normierungsstichprobe ein höheres Testergebnis erzielt haben. Bei einem Prozentrangplatz von 50 hätten genauso viele Personen ein höheres Ergebnis wie Personen ein niedrigeres Ergebnis, bei einem Prozentrangplatz von 20 zeigten 80% der Normierungsstichprobe eine höhere Testleistung.

1.3.6 Weiterführende Literatur

Krauth, J. (1995). *Testkonstruktion und Testtheorie*. Weinheim: Beltz.

Ingenkamp, K. & Lissmann, U. (2008). *Lehrbuch der pädagogischen Diagnostik* (6. Aufl.). Weinheim: Beltz.

Bühner, M. (2006). *Einführung in die Test- und Fragebogenkonstruktion* (2. Aufl.). München: Pearson Studium.

Rost, J. (2004). *Lehrbuch Testtheorie – Testkonstruktion* (2. Aufl.). Bern: Huber.

Moosbrugger, H. & Kelava, A. (Hrsg.). (2008). *Testtheorie und Fragebogenkonstruktion*. Heidelberg: Springer.

1.3.7 Verständnis- und Diskussionspunkte

1. Diskutieren Sie mögliche Ursachen, aus denen der Messfehler resultieren kann.
2. Unter welchen (ggf. auch unrealistischen) Bedingungen wäre die Reliabilität eines Tests gleich 1?
3. Was bedeutet „erwartungskonforme Korrelation“ im Hinblick auf die Validität eines Tests?
4. Welche Vorteile bietet die Odds-Even-Reliabilität (gerade-ungerade-Reliabilität) gegenüber der Split-Half-Reliabilität?
5. Diskutieren Sie, warum die Normierung eines Tests ein wichtiges Nebengütekriterium darstellt.

1.4 Inhaltlicher Anwendungsbereich/Phänomenbereich

In diesem Kapitel soll eine Auswahl an Personenmerkmalen vorgestellt werden, die sich mit Hilfe standardisierter Testverfahren erfassen lassen. Dabei nehmen wir zwei Beschränkungen vor. Zum einen beschränken wir uns auf Personenmerkmale, die für pädagogische und schulische Kontexte relevant sind. Alle Personenmerkmale, die eher im Rahmen klinischer Fragestellungen interessant sind oder aber keinen Beitrag für primär pädagogische Entscheidungen leisten, werden in diesem Kapitel nicht behandelt. Außerdem teilen wir die verschiedenen Personenmerkmale in einige wenige Klassen ein und besprechen diese Klassen dann anhand eines prototypischen Beispiels. Auf diese Weise wird die enorme Menge an Personenmerkmalen auf ein überschaubares Maß gebracht. Eine weitere Einschränkung erfolgt aus didaktischen Gründen: Wir beschreiben ausschließlich Personenmerkmale, die sich durch etablierte und standardisierte Testverfahren diagnostizieren lassen. Der Sinn dahinter ist, dass in diesem Zuge „best practice“-Beispiele präsentiert werden, die die testtheoretischen Inhalte des letzten Kapitels 1.3 auf eine gute Weise wiederholen und illustrieren. Neben diesem Wiederholungseffekt soll es in diesem Kapitel jedoch vornehmlich um folgende Fragen gehen:

- Welche Schülermerkmale sind von besonderer Bedeutung für schulische Leistung?
- Wie können Schulleistung und schulleistungsrelevante Merkmale beurteilt werden?

In diesem Kapitel wenden wir uns zwei Klassen von Personenmerkmalen zu, die im schulisch-pädagogischen Kontext von besonderer Bedeutung sind. In Anlehnung an Hosenfeld und Schrader (2006) unterscheiden wir zwischen Schulleistungsmerkmalen und (schul)leistungsrelevanten Merkmalen (Abbildung 11). Schulleistungsmerkmale lassen sich durch Schulleistungstests erfassen, die die Leistungsfähigkeit eines Schülers in einem oder mehreren Schulfächern testen. Schulleistungsrelevante Merkmale hingegen umfassen sowohl kognitive als auch motivationale Personeneigenschaften, die nicht direkt als ein Aspekt von Schulleistungsfähigkeit angesehen werden können, die jedoch eine Voraussetzung für Schulleistung sind. Bspw. ist eine angemessene Motivation eine notwendige Voraussetzung für jegliche Art von Leistung. Eine präzise und eindeutige Zuordnung einzelner schulleistungsrelevanter Merkmale in die Kategorien „kognitiv“ oder „motivational“ ist nicht immer möglich (und auch nicht notwendig). Daher sollte die Einteilung auch nicht als Dichotomie verstanden werden. Vielmehr bezeichnen wir damit die Pole einer gemeinsamen Dimension.

Eine Besprechung sämtlicher schulleistungsrelevanter Merkmale würde den Rahmen des Studienbriefs sprengen, daher behandeln wir, wie oben bereits angekündigt, exemplarisch die kognitiven und motivationalen Merkmale, die zum einen Bedingungen für die Schulleistung darstellen und zum anderen einen hohen Praxisbezug aufweisen. Wie aus Abbildung 11 hervorgeht, sind dem kognitiven Pol schulleistungsrelevanter Merkmale die Intelligenz und das Vorwissen zugeordnet. Da der Bereich der Intelligenz sehr umfassend ist, wird in diesem Kapitel ausschließlich dieser als Vertreter kognitiver Merkmale besprochen. Die Erfassung von Wissen und Vorwissen ist dann Gegenstand des Kapitels 1.5. Die Merkmale Selbstkonzept und Selbstwirksamkeitserwartung beinhalten sowohl kognitive als auch motivationale Aspekte und sind aus diesem Grund in der Mitte der Dimension angesiedelt. Im Studienbrief wird jedoch zunächst auf das Merkmal Selbstkonzept eingegangen. Anschließend wird die Diagnose der Motivation besprochen.

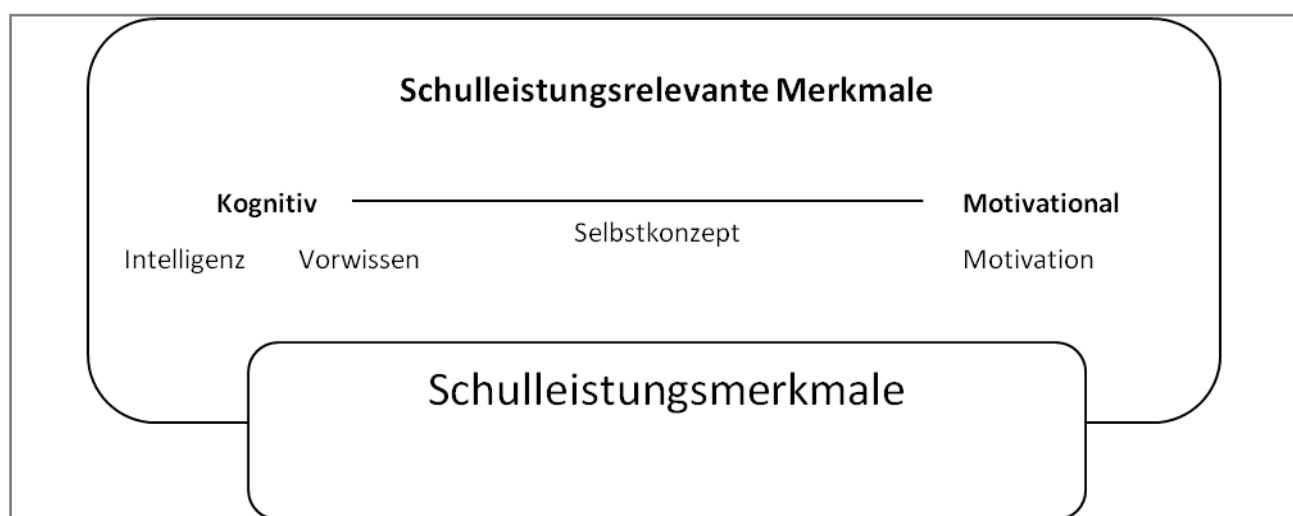


Abbildung 11: Schulisch-pädagogisch relevante Personenmerkmale

1.4.1 Schulleistungsmerkmale

Schulleistung wird im schulischen Alltag durch Zensuren quittiert. Heller (1984) definiert Schulleistung als „das gesamte Leistungsverhalten im Kontext schulischer Bildungsbemühungen.“ Rindermann und Kwiatkowski (2010) beschreiben Schulleistung als die „individuelle Leistung eines Schülers und daraus abgeleitet seine kognitiven Schulfähigkeiten (Wissen und Verständnis oder Denken und Wissen), indirekt auch Intelligenz und andere schulleistungsrelevante Personen- und Umweltmerkmale bis zu Erziehungsstilen und Bildungsorientierung der Eltern.“ Diese Definition verdeutlicht, dass Schulleistung sehr viele Facetten hat und von vielen Faktoren beeinflusst wird. Einer dieser Faktoren sind sicherlich die Kenntnisse und Fähigkeiten eines Schülers. Schulleistung ist allerdings mit Kenntnissen und Fähigkeit nicht gleichzusetzen: Die erforderliche Fähigkeit kann zwar vorhanden sein, aber im entscheidenden Moment, bspw. einer Prüfungssituation, nicht abgerufen werden. Daher unterscheidet man zwischen Performanz und Potenzial. Kenntnisse und Fähigkeiten stellen ein Potenzial dar, während Performanz als „Ausdruck oder die Anwendung von Fähigkeiten in lebensweltlich relevanten Situationen, etwa in der Schule oder allgemeiner in Ausbildungssituationen, beim Lernen, im Wissenserwerb und in Prüfungen, in der beruflichen Tätigkeit oder in Anforderungssituationen des Alltagslebens“ definiert wird (Rindermann & Kwiatkowski, 2010). Schulleistung kann im Sinne einer Performanz abgegrenzt werden von Personenmerkmalen wie z.B. Intelligenz o.ä., die im Sinne eines Potenzials zu interpretieren sind. Um diese Abgrenzung sprachlich zu untermauern, soll im vorliegenden Studienbrief zwischen „Performanz“ (Schulleistung) und „Potenzial“ (kognitive Grundfähigkeiten) unterschieden werden. Zur Erfassung schulischer Performanz stehen standardisierte Diagnoseinstrumente zur Verfügung, die als Schulleistungstests bezeichnet werden.

Potenzial vs.
Performanz

1.4.1.1 Schulleistungstests

Schulleistungstests lassen sich im Hinblick auf Messintention und Durchführungsbedingungen unterscheiden. Die Messintention eines Leistungstests kann sich auf die Erfassung allgemeiner Schulleistung oder aber auf eine bestimmte Teilleistung richten. Zudem wird zwischen bezugsgruppenorientierten und kriteriumsorientierten Schulleistungstests differenziert (vgl. Kap. 1.2). Bezugsgruppenorientierte Schulleistungstests sind Tests, bei denen das individuelle Ergebnis mit den an einer relevanten Stichprobe (meist Klassenstufe) ermittelten Ergebnissen verglichen wird. Ein kriteriumsorientierter Test dagegen ist ein wissenschaftliches Routineverfahren zur Untersuchung der Frage, ob und eventuell wie gut ein bestimmtes Lehrziel erreicht ist. Die hierbei verwendeten Testaufgaben sind nicht identisch mit dem Lehrziel, sondern repräsentieren es nur und dienen dazu, den individuellen Fähigkeitsgrad eines Schülers mit einem gewünschten Fähigkeitsgrad zu vergleichen“ (Fricke, 1973; Ingenkamp & Lissmann, 2008).

Definition: Schulleistungstests

Schulleistungstests sind Verfahren der Pädagogischen Diagnostik, mit deren Hilfe Ergebnisse geplanter und an Curricula orientierter Lernvorgänge möglichst objektiv, zuverlässig und gültig gemessen und durch Lehrende oder Beratende ausgewertet, interpretiert und für pädagogisches Handeln nutzbar gemacht werden können (Ingenkamp & Lissmann, 2008).

Unterschieden werden kann weiterhin zwischen Mehrfächertests, welche die Schulleistung in relevanten Fächern überprüfen und Tests, die auf die Erfassung der Leistung in einem bestimmten Bereich (bspw. Mathematik oder Leseverständnis) abzielen. Letztere eignen sich insbesondere zur Diagnose von Teilleistungsstörungen und werden daher im entsprechenden Abschnitt behandelt. Mehrfächertests finden ihre Anwendung in der Schullaufbahnberatung, wo sie eine datengestützte Prognose für die schulische Leistungsentwicklung ermöglichen. Darüber hinaus können Schulleistungstests zur Überprüfung des Vorwissens eingesetzt werden. Im vorliegenden Abschnitt sollen stellvertretend für die Mehrfächertests der Hamburger Schulleistungstest für 4. und 5. Klassen vorgestellt werden.

1.4.1.2 Hamburger Schulleistungstest für vierte und fünfte Klassen

Der Hamburger Schulleistungstests (HST) gehört zu den sogenannten „Mehrfächertests“ und erfasst verschiedene relevante Aspekte schulischen Lernens. Um die angestrebte Lehrzielvalidität zu gewährleisten, wurde der HST laut Testbeschreibung unter Berücksichtigung der curricularen Anforderungen der Klassen 4 und 5 konstruiert. Der Test umfasst 14 Untertests, die auf 5 Inhaltsbereiche (Subskalen) verteilt sind. Besonderes Merkmal des HST ist, dass die Informationsentnahme aus Karten, Tabellen und Diagrammen als Indikator für Schulleistung verstanden wird. Im Hinblick auf die zunehmende Relevanz selektiver Informationsentnahme, bspw. aus dem Internet, scheint die Erfassung dieser Kompetenz als sinnvoll. Da der HST sehr umfangreich ist, sollte er an zwei Tagen (eine Doppelstunde am ersten Tag und eine Einzelstunde am Folgetag) durchgeführt werden.

Diagnoseziel

Aufgabe: Datenbankrecherche

Besuchen Sie im Internet die UDiKom-Testdatenbank (<http://tests.udikom.de/>). Suchen Sie dort den Hamburger Schulleistungstest. Welche fünf Inhaltsbereiche werden durch den HST abgedeckt?

1. Subskala: _____
2. Subskala: _____
3. Subskala: _____
4. Subskala: _____
5. Subskala: _____

Validität des
HST

Hinweise auf Kriteriumsvalidität ergeben sich u.a. aus einer signifikanten Korrelation mit dem Notendurchschnitt ($r = -0,73$). Betrachtet man die Korrelation zwischen der Mathematiknote und den Ergebnissen des HST im Bereich Mathematik, ist eine etwas niedrige Validität festzustellen ($r = -0,57$). Die negative Korrelation ergibt sich aus dem Umstand, dass das Verhältnis zwischen Schulnote und Leistung nicht dem Verhältnis zwischen Testergebnis und Leistung entspricht: während eine „hohe“ (schlechte) Note wie 5 einer mangelhaften Leistung entspricht, beschreibt ein „hohes“ Testergebnis eine gute Leistung. So entsteht eine erwartungskonforme, aber negative Korrelation: je besser (höher) der Testscore, desto besser (niedriger) die Note.

Korrelation
mit Noten

1.4.1.3 Was ist ein „guter“ Schulleistungstest?

Um die Zweckmäßigkeit und die Qualität eines Schulleistungstests zu bewerten, schlägt Langfeldt (1984) folgende „Prüfsteine“ vor. Die „Prüfsteine“ beschränken sich dabei keinesfalls ausschließlich auf Schulleistungstests, sondern können auch als Bewertungskriterien für andere Testverfahren herangezogen werden.

1. Überprüft der Test das, was unterrichtet wurde?
2. Ist der Test reliabel (zuverlässig) genug?
3. Wie präzise ist ein individueller Testpunktwert?
4. Wie wird eine objektive Testdurchführung gesichert?
5. Wie wird die Auswertungsobjektivität gewährleistet?
6. Wie ist der Test normiert?
7. Gibt es Paralleltests?
8. Wie sind die Ergebnisse inhaltlich zu interpretieren?
9. Wie lange dauert der Test?
10. Wie alt ist der Test?

Qualitäts-
kriterien

Darüber hinaus bietet Langfeldt (1984) eine Orientierungshilfe zur Einschätzung der Testqualität auf Basis der angegebenen Testkennwerte. So sollten „brauchbare“ Schulleistungstests einen Validitätswert von mindestens $r = 0,60$ (wenn ein Zusammenhang und nicht eine Unabhängigkeit vermutet wurde, vgl. Kap. 1.3.4.2) und einen Reliabilitätskoeffizienten von mindestens Cronbachs $\alpha = 0,80$ aufweisen. Die Normierungsstichprobe sollte mindestens 500 Personen umfassen.

1.4.2 Schulleistungsrelevante Merkmale: Intelligenz

Intelligenz ist sicherlich eines der bedeutsamsten Personenmerkmale, die Schulleistung beeinflussen. Ihre Diagnostik erfordert allerdings vertiefte methodische Kenntnisse und sollte daher von entsprechend ausgebildeten Psychologen durchgeführt werden. Entsprechend dienen die folgenden Abschnitte nicht der Befähigung zur Durchführung von Intelligenztests. Das angestrebte Ziel ist vielmehr die Herstellung von Transparenz, um Intelligenzdiagnostik, die bspw. im Rahmen der Hochbegabendiagnostik durchgeführt wird, und ihre Ergebnisse nachvollziehen und bewerten zu können.

Intelligenz –
Was ist das?

Was Intelligenz genau ist, ist eine Frage, die in der Literatur unterschiedlich beantwortet wird und zu einer großen Anzahl unterschiedlicher Definitionen geführt hat. An dieser Stelle soll diese Diskussion um die „richtige“ Definition nicht abgebildet werden. Stattdessen wählen wir einfach eine Definition, wie sie von Amelang und Schmidt-Atzert (2006) angeboten wird und eine gleichermaßen prägnante und praxisrelevante Definition von Intelligenz darstellt.

Definition: Intelligenz

Unter Intelligenz wird das Potenzial einer Person verstanden, kognitive Leistungen zu erbringen. Eine hoch intelligente Person kann, muss aber nicht gute Leistungen in Schule und Beruf zeigen. Motivationale Gründe oder ungüns-

tige Arbeitsbedingungen können dazu führen, dass die Person nicht die Leistung erbringt, zu der sie eigentlich fähig wäre.

Das Brickenkamp-Testkompendium allein umfasst 57 verschiedene Testinstrumente zur Intelligenzmessung. Auf der Homepage des Testverlags Hogrefe sind 27 Intelligenztests für Kinder und Jugendliche aufgeführt. Die Vielzahl der verfügbaren Testinstrumente ist ein Indikator für die Popularität von Intelligenztests, welche sich z.B. durch die Fähigkeit von Intelligenztests erklären lässt, zuverlässig Niedrig- und Hochbegabung zu diagnostizieren. In der Tat gelten Intelligenztests daher als die wohl erfolgreichsten psychologischen Diagnoseverfahren (Amelang & Schmidt-Atzert, 2006). Ein wesentliches Merkmal zur Unterscheidbarkeit von Intelligenztests ist die Messintention, also die Frage, wie Intelligenz theoretisch durch ein Modell beschrieben wird (vgl. 1.3.4.2). Während einige Testinstrumente wie bspw. das Leistungsprüfsystem (LPS) ein Intelligenzmodell, das viele verschiedene Dimensionen beschreibt, durch sehr heterogene Untertests abbilden erfassen andere Testinstrumente Intelligenz mit nur einer Skala als allgemeines intellektuelles Gesamtpotenzial, welches auch als Grundintelligenz oder g-Faktor der Intelligenz bezeichnet wird. Andere Verfahren sind an der Ausprägung spezifischer Intelligenzfaktoren interessiert.

Im vorliegenden Studienbrief werden Intelligenztests anhand dreier Merkmale klassifiziert. Das Erste betrifft die Frage nach der Anzahl der *Dimensionen*, die das zu Grunde gelegte theoretische Modell der Intelligenz beschreibt. Das zweite Merkmal betrifft die *Sprachfreiheit* der Tests. Schließlich wird danach unterschieden, ob der Test unter strikten *Zeitvorgaben* durchgeführt werden muss oder nicht.

Klassifikation
von Intelli-
genztests

Bezüglich der Dimensionalität unterscheiden sich Testinstrumente dahingehend, ob der Test die allgemeine Intelligenz (general factor) „g“ oder einen oder mehrere Aspekte von Intelligenz erfasst. Es geht also darum, ob Intelligenz eindimensional als ein übergreifendes, globales Konstrukt verstanden wird oder ob verschiedene „Intelligenzen“ angenommen werden. Eindimensionale Intelligenztests sind rasch durchführbar und liefern eine globale Einschätzung des intellektuellen Potenzials. Mehrdimensionale Tests dagegen ermöglichen die Erfassung eines kognitiven Profils mit speziellen Stärken und Schwächen.

Dimensiona-
lität

Die Frage nach der Sprachfreiheit stellt sich insbesondere, wenn Schüler mit einer anderen Muttersprache getestet werden sollen. Soll bspw. die Intelligenz eines Schülers mit Migrationshintergrund überprüft werden, der die jeweilige Landessprache noch nicht beherrscht, so eignet sich ein sprachfreier Test um sprachbedingte Benachteiligung auszuschließen.

Sprachge-
bundenheit

Intelligenztests unterscheiden sich nicht nur in Bezug auf die theoretischen Merkmale, sondern auch nach den Durchführungsbedingungen, bspw. ob der Test in Einzeltestung oder in der Gruppe erfolgt. Ein Gruppentest kann aus ökonomischen Gründen sinnvoll sein. Aus motivationalen Gründen ist es jedoch teilweise ratsam, eine Einzeltestung durchzuführen. Das ist insbesondere dann der Fall, wenn es sich um Testpersonen mit kognitiven Beeinträchtigungen oder um die Diagnose besonderen Förderbedarfs handelt. Intelligenztests lassen sich zudem in sogenannte „Speedtest“ und „Powertests“ aufteilen. Bei Speedtests sind enge zeitliche Vorgaben für die Testbearbeitung gegeben, die zu einer Belastung während der Testbearbeitung führen (sollen). Bei Powertests entfällt dieser Zeitdruck. Die Entscheidung, ob der Test eine starke zeitliche Begrenzung vorgeben sollte oder nicht, sollte davon abhängig gemacht werden, ob die Zeitbegrenzung in Kombination mit Faktoren wie Leistungsängstlichkeit oder Sprachproblemen des Schülers zu verzerrten Ergebnissen führen kann. Trifft dies zu, sollte die Entscheidung zugunsten Powertests ausfallen, bei denen zwar keine eng bemessene Zeitbegrenzung vorgegeben wird, bei denen jedoch die Schwierigkeit der Items graduell ansteigt.

Zeitvorgaben

In den folgenden Abschnitten sollen einige ausgewählte Intelligenztests vorgestellt und besprochen werden.

1.4.2.1 Eindimensionale Intelligenztests: Der Hamburg-Wechsler-Intelligenz-Test

Um die Kategorie eindimensionaler Intelligenztests zu veranschaulichen, soll an dieser Stelle der Hamburg-Wechsler Intelligenztest (HAWIK) vorgestellt werden (für weiterführende Informationen siehe <http://tests.udikom.de/>). Dieser enthält zwar 13 verschiedene Untertests (Tabelle 2), weshalb man annehmen könnte, dass er ein mehrdimensionales Modell von Intelligenz abbilden soll. Aus Wechslers Definition der Intelligenz wird jedoch deutlich, dass er nicht verschiedene „Intelligenzen“ unterscheidet, sondern Intelligenz als ein globales Potenzial versteht.

Diagnoseziel

Definition: Intelligenz

Wechsler (1964) definiert Intelligenz als „globale oder zusammengesetzte Fähigkeit des Individuums, zweckvoll zu handeln, vernünftig zu denken und sich mit seiner Umgebung wirkungsvoll auseinanderzusetzen.“

Der Test spiegelt das Intelligenzmodell in Form von 13 Untertests (Subtests) wider, die sich in einen Handlungsteil und einen Verbalteil gliedern und in einer festgelegten Reihenfolge durchgeführt werden. Daraus folgt, dass Intelligenz gemäß des HAWIK als die Summe verbaler und praktischer Fähigkeiten operationalisiert wird, die durch die Untertests gemessen werden. Diese Summe ist jedoch zunächst wenig aussagekräftig (vgl. Kap.1.2.2), sondern ist in

Bezug auf eine entsprechende Altersnorm zu interpretieren. Der Test liefert differenzierte Altersnormen. Dies ist von besonderer Bedeutung für einen auf Kinder und Jugendliche ausgerichteten Intelligenztest. Der HAWIK-III wird diesem Anspruch gerecht, da er in Altersgruppen gestaffelt ist, die sich jeweils um nur 4 Monate unterscheiden. Die Größe der Stichprobe innerhalb einer Alterskohorte liegt dadurch allerdings nur zwischen 35 und 60 Personen.

Abkürzung	Untertest	Beispielaufgabe
AW	Allgemeines Wissen	In welcher Himmelsrichtung geht die Sonne unter?
GF	Gemeinsamkeiten finden	Was ist das Gemeinsame an Hemd und Schuh?
RD	Rechnerisches Denken	Franz liest 3 Seiten in 5 Minuten. Wie viele Minuten braucht er für 24 Seiten?
WT	Wortschatz-Test	Was ist ein Brot?
AV	Allgemeines Verständnis	Warum haben Autos Sicherheitsgurte?
ZN	Zahlen nachsprechen	3-4-1-7
BE	Bilder ergänzen	Was fehlt auf dem Bild? Fehlende Details benennen oder zeigen
ZS	Zahlen-Symbol-Test	Umwandlungstabelle mit Zahlen und Symbolen (z.B. +). Symbole in Felder und Zahlen eintragen.
BO	Bilder ordnen	Bilder in die richtige Reihenfolge bringen
MO	Mosaik-Test	Zweifarbige Muster mit 2, 4 bzw. 8 Klötzchen nachlegen
FL	Figurenlegen	Zerschnittene Figuren („Puzzle“) zusammenfügen
SS	Symbolsuche	Zwei Gruppen von Symbolen vorgegeben. Ankreuzen, ob ein Symbol in beiden Gruppen enthalten ist
LA	Labyrinthtest	Linie vom Zentrum zum Ausgang eines Labyrinths ziehen

Tabelle 2: Untertests des HAWIK-III

Der Vergleich des Testergebnisses mit der jeweiligen Altersnorm führt – wie bei anderen Intelligenztests auch – zur Berechnung des sogenannten „Intelligenzquotienten“.

Definition: Intelligenzquotient

Der Intelligenzquotient (IQ) zeigt das intellektuelle Leistungsvermögen einer Person im Vergleich zu einer Normstichprobe vergleichbaren Alters an. Üblicherweise wird das durchschnittliche intellektuelle Leistungsniveau einer Altersgruppe auf 100 IQ-Punkte festgelegt. Eine Standardabweichung beträgt üblicherweise 15 IQ-Punkte.

Durch-
führung
Auswertung
Interpreta-
tion

Durch die zahlreichen, recht komplexen Untertests, fordert die Durchführung des HAWIK-III viel Übung seitens des Testleiters. Die Aufgaben werden mit Hilfe einer Lösungsschablone ausgewertet, und die Rohwerte werden addiert. Die Summe ergibt die Punktezahl für den jeweiligen Untertest. Ein spezielles Programm übernimmt die Auswertung der Rohwerte, einschließlich der Ermittlung des Intelligenzquotienten und einer graphischen Darstellung der individuellen Intelligenzprofile. Zur Interpretation der Ergebnisse, insbesondere zur Erklärung schwacher Subtestergebnisse, bietet der HAWIK differenzierte Informationen. Amelang und Schmidt-Atzert (2006) sehen dies als zentralen Vorteil des Instrumentariums: „Der HAWIK-III stellt trotz einiger kleiner Unzulänglichkeiten ein brauchbares und nützliches Intelligenztestverfahren für Kinder und Jugendliche dar. Die Informationsausbeute ist groß. Der Test liefert neben dem Intelligenzquotienten viele Informationen über die Stärken und Schwächen des Probanden“.

Aufgabe: Datenbankrecherche

Der HAWIK-III liefert einen IQ-Wert, der das intellektuelle Potenzial von Kindern oder Jugendlichen ausdrücken soll. Er kann jedoch auch genutzt werden, um Stärken und Schwächen in vier verschiedenen Teilleistungsbereichen zu identifizieren. Um welche Teilleistungsbereiche handelt es sich (siehe <http://tests.udikom.de/>)?

1.4.2.2 Mehrdimensionale Intelligenztests: Das Leistungsprüfsystem

Diagnoseziel

Das Leistungsprüfsystem (LPS) spiegelt das Intelligenzmodells von Thurstone (1938/1947) wider. Dieses Modell beschreibt Intelligenz mehrdimensional, indem es 7 unabhängige sogenannte „Primärfaktoren“ der Intelligenz postuliert. Jeder Primärfaktor wird mit mindestens zwei Untertests bzw. 80 Aufgaben getestet. Das LPS eignet sich insbesondere dann, wenn eine möglichst differenzierte Aussage über verschiedene kognitive Fähigkeiten vorliegen soll.

Aufgabe: Datenbankrecherche

Suchen Sie unter <http://tests.udikom.de/> das Leistungsprüfsystem und sammeln Sie dort Informationen über die verschiedenen Untertests. Wie würden Sie die sieben Primärfaktoren benennen, die durch diese 14 Untertests abgebildet werden sollen?

1.4.2.3 Sprachfreie Tests: Der Culture Fair Test 20

Sprachfreie Tests werden vornehmlich dann eingesetzt, wenn (mutter-)sprachliche oder sozio-kulturelle Einflüsse auf die Testleistung vermieden werden sollen. Sie werden daher auch als „kulturfreie“ Tests bezeichnet. Jedoch gibt es bislang noch keinen Intelligenztest, der die Intelligenz vollkommen unabhängig von sozio-kulturellen Einflüssen messen könnte. Daher wird mittlerweile eher von „kulturfairen“ Tests gesprochen.

Sprachfreie Tests basieren auf dem Intelligenzkonzept von Cattell (1968), welches zwischen fluider und kristalliner Intelligenz unterscheidet. Die kristalline Intelligenz bezieht sich auf die „Sammlung gelernter Kenntnisse, die sich ein Mensch angeeignet hat, in dem er seine fluide Intelligenz in der Schule anwandte“ (Cattell & Piaggio, 1973). Diese kristallinen Intelligenzanteile sind entsprechend stark von sprachlichen Fähigkeiten und der Kultur abhängig. Fluide Intelligenz dagegen bezeichnet bildungsunabhängige, intellektuelle Fähigkeiten wie „die Fähigkeit komplexe Beziehungen in neuartigen Situationen wahrnehmen und erfassen zu können“ (Cattell, 1968). Sprachfreie bzw. kulturfaire Tests, wie z.B. der Culture Fair Test 20 (CFT 20), zielen daher meist auf die Erfassung fluider Intelligenzanteile ab.

Der CFT 20 ist auf die Erfassung der fluiden Intelligenz ausgerichtet und umfasst ausschließlich sprachfreie Aufgaben, die unabhängig von erlerntem Wissen und daher kulturfrei sind. Der CFT 20 setzt sich aus 4 Untertests mit insgesamt 92 Aufgaben zusammen, die es erfordern, Figurenreihen fortzusetzen, Figuren zu klassifizieren, Figurenmatrizen zu vervollständigen und topologische Schlüsse ziehen. Innerhalb eines Untertests sind die Aufgaben nach Schwierigkeit gestaffelt. Die Aufgaben sind auf zwei identische Testformen verteilt, wobei der erste Teil auch als Kurzversion eingesetzt werden kann. Der CFT 20 kann als Gruppentest aber auch in Einzeltestung mit Testpersonen im Alter von 8-70 Jahren durchgeführt werden. Jeder Untertest beginnt zunächst mit Einführungsaufgaben, um die Testpersonen mit den Anforderungen des jeweiligen Untertests vertraut zu machen. Die Durchführungszeit beträgt ca. 55 Minuten (Kurzversion 35 Minuten), was eine verhältnismäßig rasche und ökonomische Einschätzung der Grundintelligenz ermöglicht. Auch die Auswertung der Testergebnisse gestaltet sich durch eine verfügbare Auswertungsschablone als objektiv und zeitlich ökonomisch.

Diagnoseziel

Durchführung
Auswertung

1.4.3 Schulleistungsrelevante Merkmale: Motivation

Intelligenz von Schülern, wie wir sie eben besprochen haben, ist zwar eine notwendige, aber noch lange keine hinreichende Bedingung für schulischen Erfolg. Schulerfolg wird durch das Zusammenspiel verschiedener Faktoren beeinflusst. Als wichtiger Bedingungsfaktor gilt hierbei auch die Motivation. Motivation hat sich im Hinblick auf Lernen und Leistung fest im erziehungswissenschaftlichen Diskurs sowie in der Alltagssprache etabliert. Sie umfasst nach Langfeldt (2006) all diejenigen Prozesse, die zielgerichtete Verhaltensweisen in konkreten Situationen auslösen und aufrechterhalten. Der Motivation werden bestimmte lernleistungsförderliche Funktionen zugeschrieben. So geht bspw. Ormond (2006) davon aus, dass Motivation zur Verbesserung kognitiver Prozesse und zur Leistungssteigerung beiträgt. Weiterhin hat Motivation einen Einfluss darauf, was als zufriedenstellend empfunden wird und bestimmt daher auch Verhaltensabsichten und Intentionen. Bezieht sich die Motivation auf Schulleistung, so kann zwischen Lernmotivation und Leistungsmotivation unterschieden werden.

Theoretischer
Hintergrund

Definition: Lernmotivation

Lernmotivation bezeichnet die Form der Motivation, welche die Absicht oder die Bereitschaft einer Person beschreibt, sich in einer konkreten Situation mit einem Gegenstand lernend auseinander zu setzen (Wild et al., 2001).

Definition: Leistungsmotivation

Leistungsmotivation bezeichnet die Absicht, etwas zu leisten, Erfolge zu erzielen und Misserfolge zu vermeiden, wobei zur Bewertung des Erfolges bzw. Misserfolgs ein individuell verbindlicher Bewertungsmaßstab herangezogen wird (Wild et al., 2001). Auf Basis dieser Definition kann zwischen den Dimensionen „Annäherung an Erfolg“ und „Vermeidung von Misserfolg“ unterschieden werden (Langfeldt, 2006). Schüler sind motiviert durch die Absicht, Erfolge zu erreichen oder Misserfolg zu vermeiden.

1.4.3.1 Erfassung der Lern- und Leistungsmotivation: Der SELMO

Zur Diagnose der Lern-Leistungsmotivation kann bspw. auf die Skalen zur Erfassung der Lern- und Leistungsmotivation (SELMO) von Dickhäuser et al. (2002) zurückgegriffen werden. Die Skalen basieren auf der theoretischen An-

Diagnoseziel

nahme, dass die schulische Lern- und Leistungsmotivation durch verschiedene Zielorientierungen bestimmt wird. So unterscheiden Ames und Archer (1988) zwischen Lernzielen und Performanzzielen. Ein wesentliches Unterscheidungsmerkmal ist hierbei die gewählte Bezugsnorm (vgl. Kap. 1.2): Die Lernzielorientierung ist auf die Steigerung eigener Kompetenzen ausgerichtet. Der Vergleich der eigenen Leistung über einen bestimmten Zeitraum hinweg dient dem Lerner als Maßstab zur Bewertung der Kompetenzsteigerung (individuelle Bezugsnorm). Die Performanzzielorientierung dagegen bietet den Vergleich mit anderen Lernern (soziale Bezugsnorm). Lerner mit ausgeprägter Performanzzielorientierung sind weniger an einer Kompetenzsteigerung interessiert als daran, die eigene Kompetenz vor anderen Lernern zu demonstrieren bzw. einen Mangel an Kompetenz zu kaschieren. Der SELLMO spiegelt dieses theoretische Modell der Lern- und Leistungsmotivation durch vier Skalen mit insgesamt 31 Items wider (Tabelle 3).

Untertest	Beschreibung
Leistungsziele	Schüler betrachten Leistung als den Ausdruck eigener Fähigkeiten
Annäherungs-Leistungsziele:	Schüler beabsichtigen, ihre Kompetenzen vor anderen darzustellen
Vermeidungs-Leistungsziele:	Schüler beabsichtigen, Misserfolg vor anderen zu verbergen
Arbeitsvermeidung:	Schüler streben danach, möglichst wenig Anstrengung und Leistung erbringen zu müssen.

Tabelle 3: Untertests des SELLMO

Das Instrument eignet sich in besonderem Maße zur Diagnose motivationaler Defizite bei Schülern, die hinter ihrem tatsächlichen Leistungspotenzial zurückbleiben (Underachiever). Als Konsequenz können gezielte pädagogische Interventionsmaßnahmen eingeleitet werden. Hierzu zählen bspw. die Vermittlung realistischer Zielsetzungen durch die Lehrkräfte sowie die Förderung der individuellen Bezugsnormorientierung. Berger und Rockenbach (2005) beschreiben das Instrument wie folgt:

„Die SELLMO-Skalen sind ein methodisch solides und theoretisch fundiertes Instrument, um Minderleistungen bei Schülerinnen und Schülern auf Grund motivationaler Defizite aufzuklären. Hierzu sollten flankierend die intellektuellen Fähigkeiten mittels eines standardisierten Intelligenztests (z. B. HAWIK-III von Wechsler, herausgegeben von Tewes, Rossmann & Schallberger, 2001) sowie das schulische Selbstkonzept mittels des SESSKO (Schöne et al., 2002) abgeklärt werden.“

1.4.4 Schulleistungsrelevante Merkmale: Fähigkeitsselbstkonzept

Die Literatur, die sich mit dem Thema „Selbstkonzept“ auseinandersetzt, wird durch eine Vielzahl verschiedener Modelle und Vorstellungen geprägt, was zu einer „babylonischen Sprachverwirrung“ führt wie Moschner (2001) anmerkt. Laut Greve (2000) umfasst das Selbstkonzept „alle selbstbezogenen Einschätzungen, Überzeugungen und Meinungen.“ Die Definition von Shavelson, Hubner und Stanton (1976) basiert auf 7 kennzeichnenden Merkmalen, die wie folgt zusammengefasst werden können:

1. Das Selbstkonzept ist organisiert bzw. strukturiert (d.h. Personen organisieren selbstbezogene, identitätsrelevante Informationen in Kategorien und setzen diese zueinander in Beziehung).
2. Das Selbstkonzept besteht aus verschiedenen Aspekten, welche die Struktur des Selbstkonzepts einer Person widerspiegeln.
3. Das Selbstkonzept ist hierarchisch aufgebaut.
4. Das Selbstkonzept ist abnehmend stabil: die oberen Kategorien der Selbstkonzepthierarchie sind situationsunabhängig (traits), während die unteren Ebenen der Pyramide situationsabhängig sind (states).
5. Das Selbstkonzept ist Entwicklungsdynamisch: die verschiedenen Facetten werden mit zunehmendem Alter eindeutiger voneinander abgegrenzt.
6. Das Selbstkonzept hat eine deskriptive (beschreibende) und eine evaluative (bewertende) Komponente.
7. Das Selbstkonzept ist von anderen Konstrukten (z.B. Motivation) unterscheidbar.

Moschner (2001) beschreibt das Selbstkonzept als „ein mentales Modell einer Person über die eigenen Fähigkeiten und Eigenschaften.“ Nach Shavelson, Huber und Stanton (1976) enthält es neben dem emotionalen, dem sozialen und dem körperlichen Selbstkonzept eine partielle Komponente des generellen Selbstkonzepts. Zur besseren Abgrenzung der verschiedenen Konstrukte steht im vorliegenden Studienbrief lediglich das akademische bzw. schulische Fähigkeitsselbstkonzept im Vordergrund. Das Fähigkeitsselbstkonzept kann definiert werden als „Gesamtheit der kognitiven Repräsentationen eigener Fähigkeiten in akademischen Leistungssituationen“ (Dickhäuser, Schöne, Spinath & Stiensmeier-Pelster, 2002).

Dass das schulische Fähigkeitsselbstkonzept schulische Leistungen beeinflussen kann, lässt sich aus den verschiedenen Studien schließen, in denen ein signifikanter Zusammenhang zwischen der individuellen Ausprägung des Selbst-

konzepts und schulischer Leistung aufgezeigt werden konnte. Das schulische Fähigkeitsselbstkonzept klärt „nennenswerte Beiträge von (Schul-)Leistungsvarianz auf, hängt mit der Ausdauer bei der Bearbeitung von Aufgaben zusammen, beeinflusst das Wahlverhalten (z.B. in der Oberstufe), kovariiert mit Interesse und Leistungsmotivation“ (Rost et al., 2007). Die Korrelation der Schulnoten mit dem Selbstkonzept ist teilweise stärker als die Korrelation mit dem IQ (Rost et al., 2007).

Eine Metaanalyse von Hansford und Hattie (1982), die 20 Studien berücksichtigte, ermittelte eine Korrelationsstärke von $r = 0,40$. Dieser Zusammenhang ist Studien zufolge bereits während der Grundschulzeit evident: Während hochbegabte Schülerinnen und Schüler ein positives Selbstkonzept aufweisen (Rost & Hanses, 1994), konnte bei Schülerinnen und Schülern mit Leistungsschwächen ein negatives Selbstkonzept nachgewiesen werden (Hanses & Rost, 1998).

Als mögliches Erklärungsmodell für diese Befunde wird angenommen, dass das Fähigkeitsselbstkonzept leistungsschwacher Schüler durch den ständigen impliziten oder expliziten Vergleich mit stärkeren Schülerinnen und Schülern im Laufe der Schulzeit weiter absinkt. Diese Hypothese des Bezugsgruppeneffekts wird durch den Befund gestützt, dass Sonderschüler ein höheres Fähigkeitsselbstkonzept als Schülerinnen und Schüler mit gleicher Intelligenz an Regelschulen aufweisen. Der Bezugsnormorientierung (individuell, sozial, kriterial) kommt somit eine entscheidende Rolle bei der relativen Einschätzung eigener Fähigkeiten zu. Für die pädagogische Praxis würde das bedeuten, dass homogene Lerngruppen zu einem positiven Fähigkeitsselbstkonzept beitragen können, da die äquivalente Bezugsgruppe „faire“ und realistische Vergleiche ermöglicht. Dennoch muss betont werden, dass die Homogenisierung von Lernenden auch immer ein Prozess der Segregation und Ausgrenzung ist. Individuelle Förderung durch Binnendifferenzierung im Unterricht und durch gezielte Stärkung des Fähigkeitsselbstkonzepts könnten an dieser Stelle als Interventionsmaßnahmen eingesetzt werden. Die von Dickhäuser et al. (2002) vorgeschlagenen Dimensionen des schulischen Fähigkeitsselbstkonzepts finden sich in dem von ihnen entwickelten Instrument wieder, welches im folgenden Abschnitt vorgestellt werden soll.

1.4.4.1 Erfassung des schulischen Selbstkonzepts: Der SESSKO

Die Skalen zur Erfassung des schulischen Selbstkonzepts (SESSKO) umfassen mittels 22 Items vier Untertests (Tabelle 4).

Untertest	Beschreibung
Schulisches Selbstkonzept – kriterial	Einstufung der Leistungen anhand eines sachlichen Kriteriums
Schulisches Selbstkonzept – individuell	Vergleich mit den eigenen Fähigkeiten in der Vergangenheit
Schulisches Selbstkonzept – sozial	Vergleich mit anderen Personen
Schulisches Selbstkonzept – absolut	ohne Vorgabe einer Bezugsnorm erfasst

Tabelle 4: Untertests des SESSKO

Die Aufgaben innerhalb der vier Untertests erfassen gleichermaßen die Bereiche Begabung, Intelligenz, Fähigkeit, Lernfähigkeit und die Bewältigung von Anforderungen. Allerdings wurde der SESSKO dahingehend kritisiert, dass er den Schülerinnen und Schülern eine sehr ausdifferenzierte Selbsteinschätzung in den einzelnen Dimensionen abverlangt, die theoretisch als auch empirisch nicht ausreichend gestützt werden kann.

1.4.5 Weiterführende Literatur

Hosenfeld, I., & Schrader, F.W. (2006). *Schulische Leistung: Grundlagen, Bedingungen, Perspektiven*. Münster: Waxmann.

Brähler, E., Holling, H., Leutner, D. & Petermann, F. (Hrsg.). (2002). *Brickenkamp Handbuch psychologischer und pädagogischer Tests*. Band 1 und 2. Göttingen: Hogrefe.

Helmke, A. & Weinert, F. E. (1997). Bedingungsfaktoren schulischer Leistungen. In F. E. Weinert (Hrsg.), *Psychologie des Unterrichts und der Schule*. Göttingen. Hogrefe.

1.4.6 Verständnis- und Diskussionspunkte

1. In welchem Kontext bzw. bei welcher Schülerpopulation wird die Unterscheidung zwischen Fähigkeit und Leistung besonders deutlich?
2. Diskutieren Sie die Bedeutung der Aussage, Intelligenz sei etwas „Undefinierbares, aber Messbares.“
3. Suchen Sie in der UDiKom-Testdatenbank (<http://tests.udikom.de>) den Culture Fair Test (CFT) und betrachten Sie die Beschreibungen der verschiedenen Untertests. Diskutieren Sie, an welcher Stelle sprachliche oder kulturelle Einflüsse auch beim CFT eine Rolle spielen könnten.

1.5 Praktische Implikationen

In diesem Kapitel behandelte Fragen:

- *Wie lässt sich die Validität eines Leistungstests gewährleisten?*
- *Wie lässt sich die Objektivität eines Leistungstests bereits durch die Aufgabenkonstruktion erhöhen?*
- *Wie lässt sich die Reliabilität eines Leistungstests mit wenig Aufwand einschätzen?*

Im letzten Kapitel 1.4 wurde der Anwendungsbereich individualdiagnostischer Verfahren skizziert. Wir haben uns also bereits einen Überblick über die Diagnose von Schulleistungsmerkmalen und von schulleistungsrelevanten Merkmalen und deren testtheoretischen Grundlagen verschafft. Im diesem Kapitel sollen nun die praktischen Implikationen der bisher erworbenen diagnostischen Kenntnisse vorgestellt werden. Eine Möglichkeit der praktischen Umsetzung der Wissensinhalte ist die eigenständige Konstruktion eines Testinstruments, um bspw. fachspezifische Kenntnisse und Fähigkeiten zu diagnostizieren. Das Kapitel zur Testtheorie verdeutlichte zwar, dass die Erstellung eines Diagnoseverfahrens mit immensem Aufwand einhergeht, der im Schulalltag kaum zu integrieren ist. Dennoch können die erworbenen Kenntnisse auch im schulischen Alltag genutzt werden. So können bspw. bereits durch die Anwendung bestimmter Konstruktionsprinzipien bei der Entwicklung und Auswahl einzelner Testaufgaben die Voraussetzungen geschaffen werden, dass eigene Klassenarbeiten die klassischen Testgütekriterien angemessen einhalten. Klauer (1987) zeigt bspw. auf, wie durch die Art der Aufgabenkonstruktion und -auswahl die Validität eines Tests (i.S. der Inhaltsvalidität) gewährleistet werden kann, so dass auf eine empirische (und mit hohem Aufwand verbundene) Überprüfung der Validität ggf. verzichtet werden kann. Wir werden diese Vorgehensweise in diesem Kapitel kurz skizzieren.

Validität

Objektivität

Die Objektivität einer Klassenarbeit hängt von der Art der Durchführung, der Auswertung und der Interpretation des Ergebnisses ab. Wie in Kapitel 1.3.4.3 dargelegt, ist eine hohe Objektivität insbesondere durch vorab schriftlich fixierte Vorgaben und Beschreibungen zu erreichen, wie bei der Durchführung, Auswertung und Interpretation vorzugehen ist. Darüber hinaus kann die Objektivität durch eine gute Aufgabenkonstruktion erhöht werden. Wir werden in diesem Kapitel ein paar Hinweise geben, worauf bei der Konstruktion (schriftlicher) Aufgaben geachtet werden sollte.

Reliabilität

Die Einschätzung der Reliabilität erfordert die Berechnung eines Korrelationskoeffizienten. Dies ist mit heutiger Standardsoftware für Tabellenkalkulationen wie z.B. Microsoft Excel kein Problem. Auch darauf werden wir im Folgenden kurz eingehen.

1.5.1 Validität

Im schulischen Kontext ist die Validität eines Leistungstests dann gegeben, wenn der Test misst, ob Schülerinnen und Schüler das gelernt haben, was sie gelehrt bekamen und daher gelernt haben sollen. Insofern ist das Lehrziel, das eine Lehrkraft innerhalb einer Unterrichtseinheit verfolgt, die Prüfgröße für die Validität eines Leistungstests wie z.B. einer Klassenarbeit. Der erste Schritt bei der Konstruktion einer Klassenarbeit ist daher die Analyse dieses Lehrziels und eine Zerlegung des Lehrziels in Teilziele (Klauer, 2001). Die Zerlegung in Teilziele dient der Auflistung aller relevanten Inhaltsbereiche, die durch das Lehrziel angesprochen wurden. Dabei muss diese Liste zum einen vollständig sein, d.h. es darf kein Inhaltsbereich übersehen werden. Zum anderen sollte sie so feingliedrig wie möglich sein. Optimal wäre eine Liste von einzelnen Aussagen, wobei jede Aussage eine Information enthält, die die Schülerinnen und Schüler lernen sollten. Da eine derart feingliedrige Liste auf Aussagenebene jedoch äußerst aufwändig zu erstellen ist, ist eine Liste möglichst kleiner Inhaltsbereiche meist ausreichend und unter Ökonomieaspekten zu bevorzugen.

Zerlegung in
Teilziele

	Reproduzieren	Anwenden	Reflektieren	Bewerten
Inhalt A	A	B	C	D
Inhalt B	E	F	G	H
Inhalt C	I	J	K	L

Tabelle 5: Lehrzielmatrix (vgl. Klauer, 2001)

Lehrziel-
matrix

Die Inhaltsbereiche lassen sich in die Zeilen einer Tabelle eintragen (Tabelle 5). Die Spalten dieser Tabelle können dann definieren, wie die Schülerinnen und Schüler mit den Inhalten jeweils umgehen können sollen. Reicht ein einfaches Kennen und Reproduzieren der Inhalte, sollen die Inhalte auf neue Gebiete angewandt werden, soll über die Inhalte reflektiert werden oder sollen sie bewertet werden? Welches Verhalten Lernende an den jeweiligen Inhalten zeigen können sollen, ist Teil des Lehrziels und kann sich zwischen Lehrzielen verschiedener Unterrichtseinheiten entsprechend unterscheiden. Wichtig ist jedoch, dass wie bereits die Inhaltsbereiche so auch das gewünschte Verhalten möglichst umfassend beschrieben ist. Denn nur wenn sowohl die Inhaltsbereiche als auch das Verhalten voll-

ständig und umfassend definiert sind und in die Zeilen und Spalten einer Tabelle eingetragen wurden, dann repräsentieren die Zellen der Tabelle (Zellen A – L in Tabelle 5) das Lehrziel vollständig.

Die vollständige Repräsentation des Lehrziels ist eine notwendige Voraussetzung für die Inhaltsvalidität eines Tests: Für jede Zelle lassen sich Testaufgaben konstruieren, nach Möglichkeit pro Zelle dieselbe Anzahl von Aufgaben. Aus der so entstandenen Aufgabenmenge werden dann zufällig (oder stratifiziert-zufällig, s. Klauer, 1987) so viele Aufgaben ausgewählt, wie der Test am Ende enthalten soll. Ist jedoch ein Inhaltsbereich in der Tabelle nicht repräsentiert oder ist eine Verhaltensweise nicht in eine der Spalten eingetragen, dann fehlen in der Aufgabenmenge die entsprechenden Aufgaben und der Test verliert seinen Anspruch auf vollständige und umfassende Repräsentation des Lehrziels, spricht seinen Anspruch auf Inhaltsvalidität.

Auswahl von
Testaufgaben

1.5.2 Objektivität

Um die Objektivität eines Leistungstests wie einer Klassenarbeit zu gewährleisten, gilt es zum einen die Durchführungsbedingungen zu standardisieren. Dies ist im schulischen Kontext meist in hohem Ausmaß gegeben: Die Dauer ist durch den Schulstundentakt meist festgelegt, die Aufgabenstellungen werden meist in schriftlicher Form ausgeteilt, die Bedingungen werden für alle Schülerinnen und Schüler gleich gehalten, etc. Aber wie verhält es sich bspw. mit Rückfragen durch einzelne Schülerinnen und Schüler? Welche Fragen werden beantwortet und welche nicht, welche Hilfestellungen werden gegeben? Wenn ein Schüler auf eine Rückfrage einen Hinweis bekommen hat, wird dieser Hinweis dann auch allen weiteren Schülerinnen und Schülern gegeben? Wie verhält es sich mit der Kontrolle? Werden evtl. manche Schülerinnen oder Schüler stärker beobachtet als andere? Vertraut man manchen Schülerinnen und Schülern mehr als anderen? Diese Fragen stehen beispielhaft für mögliche Verletzungen der Durchführungsobjektivität. Sie sollen dazu dienen, den Blick dafür zu schärfen, wie in einer eigentlich recht standardisierten Testsituation das Objektivitätskriterium trotzdem recht leicht verletzt werden kann. Da aufgrund des trotz dieser Gefahren recht hohen Standardisierungsgrades bei Klassenarbeiten die Durchführungsobjektivität jedoch meist sehr akzeptabel ist, soll an dieser Stelle darauf nicht weiter eingegangen werden. Vielmehr wollen wir uns der Auswertungs- und Interpretationsobjektivität zuwenden, die zu einem großen Teil durch das Antwortformat der verwendeten Testaufgaben beeinflusst wird.

Durchführungs-
objektivität

1.5.2.1 Schriftliche Testaufgaben mit geschlossenem Antwortformat

Testaufgaben lassen sich zum einen danach unterscheiden, ob ihr Antwortformat schriftlich oder mündlich oder verhaltensbasiert (bspw. im Sport) ist. Wir beschränken uns im Folgenden auf Testaufgaben mit schriftlichem Antwortformat. Hier lässt sich wieder zwischen offenen und geschlossenen Antwortformaten unterscheiden. Bei offenen Antwortformaten wie z.B. kurzen Aufsätzen oder Portfolios kann die richtige oder optimale Lösung der Aufgabe in vielen verschiedenen Varianten niedergeschrieben werden bzw. es kann mehrere richtige oder optimale Lösungen der Aufgabe geben. Bei geschlossenen Aufgabenformaten wie z.B. Ergänzungsaufgaben oder Mehrfach-Wahlaufgaben (Multiple-Choice-Aufgaben) gibt es genau eine vorab bestimmte richtige Lösung, die sich durch ein genau definiertes Wort oder auch nur ein richtig gesetztes Kreuz o.ä. ausdrückt. Damit erlauben geschlossene Antwortformate deutlich weniger Spielraum bei der Auswertung des Tests, wodurch die Objektivität von Testaufgaben mit geschlossenem Antwortformat als deutlich höher einzuschätzen ist als bei Testaufgaben mit offenem Aufgabenformat.

Doch der Vorteil der höheren Objektivität von Testaufgaben mit geschlossenem Aufgabenformat geht einher mit dem Nachteil, dass diese Art der Testaufgaben meist schwieriger zu konstruieren ist. Zudem haben Testaufgaben mit geschlossenem Aufgabenformat den Ruf, nur kognitiv wenig anspruchsvolle Fähigkeiten wie den reinen Abruf von Wissen testen zu können. Dass dieser Ruf jedoch nicht gerechtfertigt ist, demonstriert u.a. Klauer (2001) eindrucksvoll anhand von Mehrfach-Wahlaufgaben aus der TIMS-Studie oder dem Mediziner-Test. Doch wenn Testaufgaben mit geschlossenem Antwortformat besser sein sollen als ihr Ruf, dann müssen bei der Aufgabenkonstruktion einige Faustregeln beachtet werden, was mit einem gewissen Aufwand verbunden ist. Diese Faustregeln sollen im Folgenden für die gängigsten geschlossenen Antwortformate besprochen werden.

Ergänzungsaufgaben

Ergänzungsaufgaben eignen sich nicht nur für die Wissensabfrage, sondern auch für (wenig komplexe) Formen der Wissensanwendung (bspw. das Lösen von Bruchrechenaufgaben) und des Verständnisses (bspw. Steigerung eines Adjektivs). Einige Beispiele (gut sowie schlecht konstruierter) Ergänzungsaufgaben finden sich in Tabelle 6.

Ergänzungsaufgaben werden konstruiert, indem aus einer Aussage ein Wort oder eine Zahl entfernt und durch eine Leerstelle ersetzt wird. Wurde das Lehrziel bis auf Aussagenebene in Teilziele zerlegt, können diese Aussagen herangezogen werden. Andernfalls müssen Aussagen konstruiert werden, die jeweils ein Teilziel repräsentieren. Wichtig dabei ist, dass die Aussagen keine wörtlichen Zitate aus Lehrtexten darstellen, da dieses ein reines Auswendiglernen seitens der Schülerinnen und Schüler befördert. Durch eine entsprechende klare Instruktion werden die Schülerinnen und Schüler dann gebeten, in die Leerstelle das fehlende Wort bzw. die fehlende Zahl einzutragen.

Faustregeln

Die Aussagen sollten so konstruiert sein, dass die Leerstellen möglichst weit am Ende des Satzes stehen. Das ermöglicht den Schülerinnen und Schülern, möglichst viele Informationen zunächst zu lesen, bevor sie eine Antwort finden müssen. Um keine Hinweise auf das Lösungswort zu geben, sollten alle Leerstellen im Text dieselbe Länge aufweisen. Zudem sollte vermieden werden, dass grammatikalische Hinweise wie z.B. Artikel manche (falsche) Lösungswörter ausschließen. Wichtig ist, dass es für jede Leerstelle genau ein einziges richtiges Lösungswort bzw. genau eine richtige Zahl gibt. Genauso sollte in einer Aussage möglichst nur eine einzige Leerstelle gesetzt werden, da ansonsten die Gefahr besteht, dass ein falsches Ausfüllen der einen Leerstelle auch zu einem fehlerhaften Ausfüllen der weiteren Leerstelle führt, sprich die Leerstellen nicht unabhängig voneinander ausgefüllt werden können.

Ergänze jede Leerstelle so, dass die Aussage stimmt. Schreibe dabei deutlich und richtig. Jede richtige Antwort gibt einen Punkt.	
Die Evolutionstheorie von _____ basiert auf dem Prinzip der _____.	Mehrere Leerstellen; Grammatikalischer Hinweis
Columbus entdeckte Amerika _____.	Mehrere mögliche richtige Antworten
$16 + 7 * 2 = \underline{\hspace{1cm}}$.	gut
In welchem Jahr wurde Helmut Kohl zum ersten Mal Bundeskanzler der Bundesrepublik Deutschland? _____	gut
Zu _____ Leerstellen frustrieren sowohl _____ als auch _____.	Mehrere Leerstellen; Leerstellen unterschiedlicher Länge; Leerstelle am Beginn der Aussage

Tabelle 6: Beispiele für Ergänzungsaufgaben inklusive Bewertungen

Zuordnungsaufgaben

Zuordnungsaufgaben bestehen aus einer Liste von Aussagen und einer Liste von Optionen. Die Aufgabe besteht darin, jeder Aussage genau eine der Optionen zuzuordnen. Zwei Beispiele für Zuordnungsaufgaben sind in Tabelle 7 dargestellt.

Aufgabe 1: Ordne jeder Persönlichkeit eine Aussage zu.		
1. Gott ist tot	1. Marx	Keine Homogenität
2. Gott als Projektion	2. Freud	Hohe Ratewahrscheinlichkeit
3. Religion als Opium des Volkes	3. Nietzsche	Keine Ordnung
4. Religion als kollektive Zwangsnervose	4. Lennon	Keine Eindeutigkeit
5. Gott ist ein Konzept	5. Feuerbach	Mangelhafte Instruktion
Aufgabe 2: In der linken Spalte steht, was eine Person erfunden hat, in der rechten Spalte stehen berühmte Erfinder. Ordne den Erfindungen ihren Erfinder zu, indem du den entsprechenden Buchstaben auf die Linie vor der Erfindung schreibst.		
___ 1. Er hat die Entkörnungsmaschine für Baumwolle erfunden.	a. Alexander G. Bell	gut
___ 2. Eine seiner Erfindungen war das Telefon.	b. Henry Bessemer	
___ 3. Er hat das Radio erfunden.	c. Thomas Edison	
	d. Guglielmo Marconi	
	e. Eli Whitney	
	f. Orville Wright	

Tabelle 7: Beispiele für Zuordnungsaufgaben inklusive Bewertungen

Faustregeln

Die Instruktionen enthalten die Angabe, ob jede Option genau einmal oder mehrfach zugeordnet werden muss und auf welche Weise zugeordnet werden soll. Die Aussagen sowie die Optionen stammen aus homogenen Inhaltslisten (nicht wie im oberen Beispiel in Tabelle 7, wo in den Optionen Politiker und Komiker vermischt werden), damit einzelne Optionen nicht mehr ins Auge stechen als andere. Es sollten nicht mehr als zehn Aussagen bzw. Optionen pro Aufgabe verwendet werden, wobei die Anzahl der Optionen größer sein sollte als die Anzahl der Aussagen, was die Ratewahrscheinlichkeit bei der Zuordnung verringert. Jeder Aussage sollte genau eine richtige Option zugeordnet werden können, wenngleich Mehrfachzuordnungen durchaus möglich sind. Darauf muss dann in der Instruktion jedoch explizit hingewiesen werden. Der Übersichtlichkeit wegen sollte eine Aufgabe inklusive der Aussagen und Optionen nicht über zwei Seiten hinweg präsentiert werden. Aussagen und Optionen sollten nummeriert sein, jedoch mit unterschiedlichen Nummerierungen (bspw. Nummern für die Aussagen und Buchstaben für die Optionen).

Wahr-/Falsch-Aufgaben

Wahr-/Falsch-Aufgaben ermöglichen eine effiziente Abdeckung umfangreicher Inhaltsbereiche, insbesondere wenn das Lehrziel und seine Teilziele bereits auf Aussagenebene definiert sind. Der Aufwand bei der Konstruktion und Auswertung der Aufgaben ist verglichen mit alternativen Antwortformaten gering. Werden bei der Konstruktion von Wahr-/Falsch-Aufgaben jedoch bestimmte Gestaltungsregeln außer Acht gelassen, reduziert sich der diagnostische Wert dieser Aufgaben auf eine reine Abfrage trivialen Faktenwissens, die mit sehr hoher Ratewahrscheinlichkeit einhergeht und Schülerinnen und Schüler zur unreflektierten Akzeptanz vereinfachter Aussagen verleiten kann. Beispiele für Wahr-/Falsch-Aufgaben sind in Tabelle 8 dargestellt.

Gib für jede Aussage an, ob sie wahr oder falsch ist.			
	wahr	falsch	
Der Monoghalea fließt nach Norden, wo er sich bei Columbus mit dem Allegheny vereint und damit den Ohio bildet.	<input type="checkbox"/>	<input type="checkbox"/>	Mehrere Aussagen
Lange Tests sind immer reliabler als kurze Tests.	<input type="checkbox"/>	<input type="checkbox"/>	absolute Formulierung
Lange Tests sind meistens reliabler als kurze Tests.	<input type="checkbox"/>	<input type="checkbox"/>	abschwächende Formulierung
Gebete sollten in der Schule verboten sein.	<input type="checkbox"/>	<input type="checkbox"/>	Meinung ohne konkreten Bezug
$5 + 3 * 2 = 16$	<input type="checkbox"/>	<input type="checkbox"/>	gut
Wenn ein Flugzeug genau auf der Grenze zwischen Deutschland und Frankreich abstürzt, wird die eine Hälfte der Überlebenden in Deutschland und die andere Hälfte in Frankreich beigesetzt.	<input type="checkbox"/>	<input type="checkbox"/>	Fangfrage

Tabelle 8: Beispiele für Wahr-/Falsch-Aufgaben inklusive Bewertungen

Bei der Konstruktion von Wahr-/Falsch-Aufgaben sollte in einer Aufgabe genau eine Aussage repräsentiert sein (und nicht mehrere Aussagen miteinander verknüpft werden wie im ersten Beispiel in Tabelle 8) und diese sollte zweifelsfrei und ohne weitere Erläuterung als wahr oder falsch bewertbar sein. Die Aufgaben sollten auf eine Wissens- und Verständnisabfrage abzielen und keine Meinungen erfragen, die in diesem Antwortformat nicht begründbar sind. Die Aussagen sollten auch typische Fehlkonzpte repräsentieren, die dann als falsch zu bewerten wären. Bei der Konstruktion von Aufgaben, die als wahr zu bewerten sind, tendiert man häufig dazu, die Aussage sehr präzise zu formulieren und damit die Aussagenlänge zu erhöhen. Diese Tendenz ist bei Aussagen, die als falsch zu bewerten sind, nicht so stark ausgeprägt, wodurch es leicht passiert, dass Wahr-Aussagen lang und Falsch-Aussagen kurz formuliert sind. Ein solcher systematischer Unterschied in der Aussagenlänge sollte unbedingt vermieden werden. Ebenso zu vermeiden sind doppelte Verneinungen, die schnell überlesen werden, oder auch absolute oder abschwächende Formulierungen, da absolut formulierte Aussagen mit hoher Wahrscheinlichkeit falsch sind, Aussagen mit abschwächenden Aussagen mit hoher Wahrscheinlichkeit wahr. Testerfahrene Schülerinnen und Schüler können davon profitieren ohne Kenntnisse und Fähigkeiten im eigentlich zu testenden Inhaltsbereich zu haben. Wie für alle Aufgaben mit geschlossenem Aufgabenformat gilt auch für Wahr-/Falsch-Aufgaben, dass die Aussagen keine direkten Zitate aus Lehrbüchern sein sollten (um ein Auswendiglernen zu vermeiden), dass keine Fangfragen zu verwenden sind und dass bei mehreren zu bewertenden Aussagen keine Systematik in der Reihenfolge von wahren und falschen Aussagen erkennbar sein sollte.

Faustregeln

Mehrfach-Wahlaufgaben (Multiple Choice)

Mehrfach-Wahlaufgaben (Multiple Choice-Aufgaben, MC-Aufgaben) bieten nicht nur die Möglichkeit einer vertiefenden Wissensabfrage, sondern auch der Erfassung von Verständnis-, Anwendungs- und Transferleistungen. Allerdings ist dafür die Konstruktion von MC-Aufgaben äußerst aufwändig. Eine MC-Aufgabe setzt sich aus einem Aufgabenstamm inklusive der Fragestellung, einer Anweisung und den Optionen zusammen (Tabelle 9).

Der Aufgabenstamm wird vollständig vor den Alternativen präsentiert. Er enthält alle notwendigen Informationen, die die Schülerinnen und Schüler für das Verständnis der Aufgabe benötigen, und nicht mehr. Auf irrelevante Ausschmückungen sollte verzichtet werden und das Vokabular sowie die Satzstruktur sollten möglichst einfach sein, um den Leseaufwand nicht unnötig zu erhöhen. Negative Formulierungen sollten vermieden werden. Zudem sollten auch bei MC-Aufgaben direkte Zitate aus Lehrbüchern vermieden werden. Der Aufgabenstamm endet mit einer direkten Frage, zu der die folgenden Optionen jeweils eine mögliche Antwort darstellen. Hierbei sollte überprüft werden, ob die Frage womöglich grammatikalische Hinweise auf die richtige Option (den Attraktor) enthält. Auf eine Abfrage persönlicher Meinungen sollte verzichtet werden.

Faustregeln
Aufgabenstamm

Die Anweisung enthält die Angabe, ob genau eine Option richtig ist oder ob keine oder mehrere Optionen richtig sein können.

Faustregeln
Optionen

Die Konstruktion der Optionen ist das Kernstück der MC-Aufgabenkonstruktion. Üblicherweise werden pro Aufgabe 3 bis 5 Optionen gegeben, die aus einer homogenen Inhaltsliste stammen und ungefähr dieselbe Länge haben sollten. Sollten alle Optionen mit denselben Worten beginnen, können diese über die Optionen geschrieben werden und mit [...] mit den Optionen verbunden werden, um den Leseaufwand zu reduzieren. Auch bei der Formulierung der Optionen gilt, dass keine direkten Zitate aus Lehrbüchern und keine absoluten oder abschwächenden Formulierungen verwendet werden sollten. Die Optionen müssen unabhängig voneinander bewertbar sein, was auch für Optionen unterschiedlicher Aufgaben desselben Tests gilt.

Bei den Optionen unterscheidet man zwischen dem Attraktor und den Distraktoren. Der Attraktor ist die richtige Option, die Distraktoren repräsentieren falsche Antworten auf die im Aufgabenstamm präsentierte Frage. Der Attraktor muss ohne weitere Erläuterung als richtig bewertbar sein. Bei mehreren MC-Aufgaben in einem Test sollte darauf geachtet werden, dass die Position der Attraktoren keine Systematik aufweist (z.B. immer die erste Option als Attraktor). Die Konstruktion der Distraktoren ist die größte Herausforderung. Sie haben die Funktion, Personen mit geringen Kenntnissen und Fähigkeiten von dem Attraktor abzulenken. Dafür müssen sie (für diese Personen) plausibel sein. Daher bieten sich insbesondere typische Fehlkonzepte als Grundlage für die Distraktorenkonstruktion an. Je ähnlicher Distraktoren zum Attraktor sind, desto schwieriger machen sie die Aufgabe.

Welcher Autor schrieb den Montageroman „Die neuen Leiden des jungen W.“?		Aufgabenstamm
Kreuzen Sie die eine richtige Antwort an.		Anweisung
<input type="checkbox"/>	Büchner	Optionen:
<input type="checkbox"/>	Fontane	Distraktor
<input type="checkbox"/>	Goethe	Distraktor
<input type="checkbox"/>	Plenzdorf	Lockvogel
<input type="checkbox"/>	Schiller	Attraktor
		Distraktor

Tabelle 9: Beispiele für eine Mehrfach-Wahlaufgaben

Faustregeln
Aufgaben-
stellung

1.5.2.2 Schriftliche Testaufgaben mit offenem Antwortformat

Aufgaben mit einem offenen Antwortformat wie z.B. Aufsätze bieten den Vorteil, bei relativ geringem Konstruktionsaufwand relativ komplexe Inhaltsbereiche und Leistungen adressieren zu können. Der geringe Aufwand bei der Konstruktion der Aufgaben wird jedoch erkauft durch einen recht hohen Aufwand bei der Auswertung (und Interpretation) der Antworten sowie durch eine häufig geringe Objektivität. Auch wenn dieses Problem wohl nie vollständig gelöst werden kann, so gibt es doch einige Maßnahmen, um zumindest die Objektivität dieser Aufgaben zu erhöhen. Diese Maßnahmen betreffen zum einen die Aufgabenstellung selbst sowie das Vorgehen bei der Auswertung.

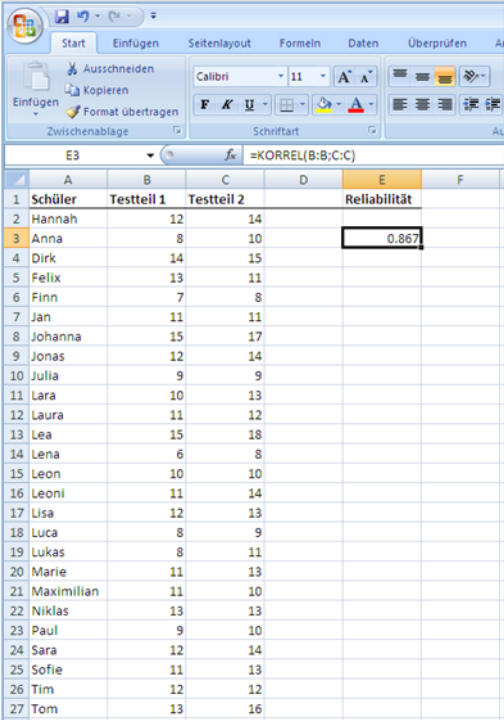
Bevor eine Aufgabenstellung formuliert wird, sollte man sich darüber bewusst sein, welche Art von Verhalten von den Schülerinnen und Schülern erwartet wird. Wenn „nur“ ein einfaches Reproduzieren oder Anwenden verlangt ist, dann lässt sich dieses auch mit Aufgaben mit geschlossenem Antwortformat erfassen. Nur wenn die Anforderungen in Richtung Bewerten oder Reflektieren gehen, ein Anforderungsniveau, für das die Konstruktion von Aufgaben mit geschlossenem Antwortformat sehr aufwändig ist, dann lohnt es sich, das offene Aufgabenformat und den damit verbundenen Auswertungs- und Interpretationsaufwand zu wählen. Die Aufgabenstellung sollte in dem Fall klar formuliert sein und keine Was-/Wer-/Wann-Fragen enthalten, die eine reine Wissensreproduktion verlangen. Sie sollte Angaben über den erwarteten Inhalt und das erwartete Verhalten/das erwartete Niveau enthalten genauso wie Angaben über evtl. Seiten- oder Zeitbeschränkungen, über die Bewertung der Organisation und Struktur des Textes sowie über den Umgang mit Rechtschreib- und Grammatikfehlern. Wenn die Aufgabenstellung das Darlegen eines Standpunktes bzw. einer Meinung verlangt, dann sollte deutlich gemacht werden, dass nicht die Meinung selbst, sondern die Begründung der Meinung bewertet wird. Aus Gründen der Vergleichbarkeit sollte auf verschiedene zur Wahl stehende Aufgabenstellungen verzichtet werden.

Faustregeln
Auswertung

Für die Auswertung sollte aus Gründen der Objektivität auf ein vorab erstelltes Bewertungsschema zurückgegriffen werden, das bspw. in Form einer Checkliste formuliert sein kann. Dieses Bewertungsschema enthält die Kriterien, nach denen die offenen Antworten zu bewerten sind (und die in der Aufgabenstellung den Schülerinnen und Schülern auch genannt wurden). Diese Kriterien werden zudem in ihrer Ausprägung beschrieben, die erreicht werden muss, um für dieses Kriterium die volle Punktzahl zu erhalten. Möglich und gängig sind auch Einschätzungen der Qualität einzelner Kriterien auf einer sog. Likert-Skala. Dabei werden verschiedene Aussagen dahingehend bewertet, inwiefern sie für den Aufsatz zutreffen. Bspw. könnte die Aussage „In dem Aufsatz werden die relevanten Pro- und Contra-Argumente klar verständlich erläutert“ dahingehend bewertet werden ob sie (1) „nicht zutrifft“, (2) „eher nicht zutrifft“, (3) „weder zutrifft noch nicht zutrifft“, (4) „eher zutrifft“ oder (5) „zutrifft“.

1.5.3 Reliabilität

Um die Reliabilität eines Tests zu erhöhen und sie zudem überprüfbar zu machen, gibt es eine goldene Regel: Je mehr Aufgaben ein Test zur Erfassung einer Leistung bzw. Fähigkeit enthält, desto besser. Ein Testergebnis, das als Summe oder Mittelwert über viele Testaufgaben berechnet wird, hat mit hoher Wahrscheinlichkeit eine höhere Reliabilität als ein Ergebnis, das nur auf sehr wenigen Testaufgaben, sprich auf sehr wenigen Messungen beruht. Je weniger Messungen, desto stärker fallen die Fehler einer Messung ins Gewicht. Um also die Reliabilität eines Tests zu erhöhen, sollte man versuchen, Tests und Klassenarbeiten zu konstruieren, bei denen dieselbe Fähigkeit wiederholt durch mehrere/viele Aufgaben getestet wird.



	A	B	C	D	E	F
1	Schüler	Testteil 1	Testteil 2		Reliabilität	
2	Hannah	12	14			
3	Anna	8	10		0.867	
4	Dirk	14	15			
5	Felix	13	11			
6	Finn	7	8			
7	Jan	11	11			
8	Johanna	15	17			
9	Jonas	12	14			
10	Julia	9	9			
11	Lara	10	13			
12	Laura	11	12			
13	Lea	15	18			
14	Lena	6	8			
15	Leon	10	10			
16	Leoni	11	14			
17	Lisa	12	13			
18	Luca	8	9			
19	Lukas	8	11			
20	Marie	11	13			
21	Maximilian	11	10			
22	Niklas	13	13			
23	Paul	9	10			
24	Sara	12	14			
25	Sofie	11	13			
26	Tim	12	12			
27	Tom	13	16			

Abbildung 12: Reliabilitätsberechnung mit Excel

Eine Mehrzahl von Aufgaben ist gleichzeitig die notwendige Voraussetzung für die Überprüfung der Reliabilität. Wie in Kapitel 1.3.4.1 dargelegt, erfolgt eine Reliabilitätsprüfung prinzipiell durch die Berechnung eines Korrelationskoeffizienten, wofür für jede getestete Person mindestens zwei Messungen vorliegen müssen. Die Berechnung einer Korrelation ist mit gängiger Tabellenkalkulationssoftware wie Microsoft Excel einfach zu realisieren. Alles, was benötigt wird, sind zwei Messungen (zwei Zahlen) für jede getestete Person, die man in ein Tabellenblatt einträgt (Abbildung 7). Diese zwei Zahlen erhält man bspw., wenn man nach der split-half- oder der odds/even-Methode (s. Kap. 1.3.4.1) die Aufgaben eines Tests in zwei Gruppen aufteilt und für jede Person eine Summe oder einen Mittelwert für diese beiden Aufgabengruppen (=Testteile) berechnet. Diese Werte sind in das Tabellenblatt so einzutragen, dass in einer Spalte (in Abbildung 12 ist das Spalte B) die Werte für den einen Testteil und in einer zweiten Spalte (Spalte C in Abbildung 12) die Werte für den anderen Testteil stehen. Excel berechnet die Korrelation der beiden Testteile, wenn man in eine leere Zelle den Befehl „=KORREL(B:B;C:C)“ schreibt, wobei „B:B“ für die Spalte mit den Werten des ersten Testteils steht und „C:C“ für die Spalte mit den Werten des zweiten Testteils. Ein Korrelationskoeffizient von $r = 0,80$ oder höher ist sehr zufriedenstellend. Für die schulische Praxis sind aber auch Koeffizienten größer $r = 0,60$ durchaus zufriedenstellend.

Berechnung
der Reliabili-
tät

1.5.4 Weiterführende Literatur

- Hanna, G.S. & Dettmer, P.A. (2004). *Assessment of effective teaching. Using context-adaptive planning*. Boston: Pearson.
- Kubiszyn, T. & Borich, G. (2003). *Educational testing and measurement. Classroom application and practice*. (7th ed.). New York: Wiley.
- Nitko, A.J. (2004). *Educational assessment of students*. (4th ed.). Upper Saddle River, NJ: Pearson.

1.6 Literatur

- Amelang, M., & Schmidt-Atzert L. (2006). *Psychologische Diagnostik und Intervention*. Heidelberg: Springer.
- Ames, C., & Archer, J. (1988). Achievement goals in the classroom: Students' learning strategies and motivation processes. *Journal of Educational Psychology*, 80, 260-267.
- Berger, U., & Rockenbach, K. (2005). Testbesprechung: Skalen zur Erfassung der Lern- und Leistungsmotivation (SELLMO). *Diagnostica*, 51, 207-211.
- Bortz, J. (2005). *Statistik für Human- und Sozialwissenschaftler*. Heidelberg: Springer.
- Bortz, J., & Döring, N. (2006). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler*. Heidelberg: Springer.
- Cattell, R. B. (1968). Are IQ-Tests intelligent? *Psychology Today*, 2, 56-62.
- Cattell, R. B., & Piaggio, L. (1973). *Die empirische Erforschung der Persönlichkeit*. Weinheim: Beltz.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297-334.
- Degen, R. (2000). *Lexikon der Psychoirrtümer. Warum der Mensch sich nicht therapieren, erziehen und beeinflussen lässt*. Frankfurt: Eichborn.
- Dickhäuser, O. Schöne, C., Spinath, B. und Stiensmeier-Pelster, J. (2002). Skalen zum akademischen Selbstkonzept: Konstruktion und Überprüfung eines neuen Instruments. *Zeitschrift für Differentielle und Diagnostische Psychologie*, 23, 393-405.
- Greve, W. (2000). Die Psychologie des Selbst: Konturen eines Forschungsthemas. In W. Greve (Hrsg.), *Die Psychologie des Selbst* (S. 15-36). Weinheim: PVU.
- Gröschke, D. (2005). *Psychologische Grundlagen für Sozial- und Heilpädagogik. Ein Lehrbuch zur Orientierung für Heil-, Sonder- und Sozialpädagogen*. Bad Heilbrunn: Klinkhardt.
- Hansford, B. C. & Hattie, J. A. (1982). The relationship between self and achievement/ performance measures. *Review of Educational Research*, 52, 123-142.
- Hanses, P. & Rost, D. H. (1998). Das "Drama" der hochbegabten Underachiever – „Gewöhnliche“ oder „außergewöhnliche“ Underachiever? *Zeitschrift für Pädagogische Psychologie*, 12, 53-71.
- Heller, K. A. (1984). Schulleistungsdiagnostik: Einleitung und Übersichtsreferat. In K. A. Heller (Hrsg.), *Leistungsdiagnostik in der Schule* (S. 15-38). Bern: Huber.
- Hosenfeld, I., & Schrader, F. W. (2006). *Schulische Leistung: Grundlagen, Bedingungen, Perspektiven*. Münster: Waxmann.
- Ingenkamp, K., & Lissmann, U. (2008). *Lehrbuch der pädagogischen Diagnostik* (6., neu ausgestattete Aufl.). Beltz Pädagogik. Weinheim: Beltz.
- Klauer, K. J. (1987). *Kriteriumsorientierte Tests*. Göttingen: Hogrefe.
- Klauer, K. J. (2001). Wie misst man Schulleistungen?. In F. E. Weinert (Hg.), *Leistungsmessungen in Schulen* (S. 103-115). Weinheim: Beltz.
- Kliemann, S. (Hrsg.) (2008). Diagnostizieren und Fördern in der Sekundarstufe I. Schülerkompetenzen erkennen, unterstützen und ausbauen. Berlin: Cornelsen Scriptor.
- Langfeldt, H.-P. (1984). Die Klassische Testtheorie als Grundlage normorientierter (standardisierter) Schulleistungstests. In: K. A. Heller (Hrsg.), *Leistungsdiagnostik in der Schule* (S. 65-98). Bern: Huber.
- Langfeldt, H.-P. (2006). *Psychologie für die Schule*. Weinheim: Beltz.
- Lienert, G. A. (1961). *Testaufbau und Testanalyse*. Weinheim: Beltz.
- Moosbrugger, H., & Kelava, A. (Hrsg.). (2007). *Testtheorie und Fragebogenkonstruktion*. Heidelberg: Springer.
- Moschner, B. (2001). Selbstkonzept. In: D. Rost (2001). *Handwörterbuch pädagogische Psychologie*, Weinheim: Beltz.
- Rost, J. (2004). *Lehrbuch Testtheorie – Testkonstruktion* (2. Aufl.). Bern: Huber.
- Rheinberg, F. (1980). *Leistungsbewertung und Lernmotivation*. Göttingen: Hogrefe.
- Rheinberg, F. (2001). Bezugsnormen und schulische Leistungsmessung. In: F. E. Weinert (Hrsg.), *Leistungsmessungen in Schulen* (S. 59-71). Weinheim: Beltz.
- Rindermann, H. & Kwiatowski, (2010). Diagnostik von Intelligenz. In C. Quaiser-Pohl & H. Rindermann (Hrsg.), *Entwicklungsdiagnostik*. München: Reinhardt.
- Rost, D.H. & Hanses, P. (1994). Besonders begabt: besonders glücklich, besonders zufrieden? Zum Selbstkonzept hoch- und durchschnittlich begabter Kinder. *Zeitschrift für Psychologie*, 202, 379-403.
- Rost, D. H., Sparfeldt, J. & Schilling, S. R. (2007). *DISK-Gitter mit SKSLF-8. Differentielles Schulisches Selbstkonzept-Gitter mit Skala zur Erfassung des Selbstkonzepts schulischer Leistungen und Fähigkeiten*. Göttingen: Hogrefe.

- Shavelson, R.J., Huber, J.J. & Stanton, G.C. (1976). *Self-concept: Validation of construct interpretations*. Review of Educational Research, 46, pp. 407-441.
- Schrader, F.-W. & Helmke, A. (2001). Alltägliche Leistungsbeurteilung durch Lehrer. In F. E. Weinert (Hrsg.) *Leistungsmessungen in Schulen* (S. 45-58). Weinheim: Beltz.
- Wild, E., Hofer, M., & Pekrun, R. (2001). Psychologie des Lerners. In A. Krapp, B. Weidenmann (Hrsg.), *Pädagogische Psychologie. Ein Lehrbuch* (S. 207-270). Weinheim: Beltz PVU.